TOWARDS AN INTEGRATIVE STUDY OF SELF

by

JORDAN LEIGHA LIVINGSTON

A DISSERTATION

Presented to the Department of Psychology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2018

DISSERTATION APPROVAL PAGE

Student: Jordan Leigha Livingston

Title: Towards an Integrative Study of Self

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Psychology by:

Elliot T. Berkman	Chairperson
Robert S. Chavez	Core Member
Jennifer H. Pfeifer	Core Member
Nicolae Morar	Institutional Representative

and

Janet Woodruff-Borden	Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded September, 2018

DISSERTATION ABSTRACT

Jordan Leigha Livingston

Doctor of Philosophy

Department of Psychology

September 2018

Title: Toward an Integrative Study of Self

The study of self within psychology has been limited in a number of ways. Two sets of empirical studies extended the study of self beyond traditional trait-based self-perception. In the first set of studies, seven hundred and eighty-nine adults listed their multiple "self-aspects" that represent meaningful elements of their lives and completed trait ratings for each of their self-aspects. The similarity between trait responses for the different self-aspects indicated the degree of "self-complexity" for a participant, as well as the degree of "self-integration." Results replicated previous findings indicating that lower self-complexity is associated with higher well-being, and that network-based approaches for measuring self-complexity were more strongly with well-being. Finally, participants who completed the same task 3 weeks later demonstrated an increase in self-integration. Broadly, the results demonstrate that network-based approaches are an effective metric for studying the structure of the self and that future work may have success using networks to inform identity-based interventions.

In the second set of studies, five hundred and ninety-four adults completed studies about personal identity and morality. Participants imagined that some trait about someone had changed and were asked to indicate the degree to which the trait change would change the person's identity. Comparisons of interest examined the degree to which

moral trait changes led to more perceived identity change than non-moral trait changes and the degree to which imagining changes to oneself versus to another person yielded differences in perceived identity change. Results replicated previous work indicating that morals lead to most perceived identity change and find that changes to self yielded large perceived identity change than changes to a friend. Moreover, neuroimaging work revealed that thinking about identity change for both targets recruits regions of the cortical midline and that thinking about moral trait words does not recruit any regions compared to thinking about non-moral trait words, challenging previous assumptions about the nature of self-perception and personal identity. Results from both sets of studies were integrated with philosophical and translational perspectives to consider the overall contributions to real-world, self-control issues and broader questions about the nature of the self.

CURRICULUM VITAE

NAME OF AUTHOR: Jordan Leigha Livingston

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Washington University in St. Louis, St. Louis, MO

DEGREES AWARDED:

Doctor of Philosophy, Psychology, 2018, University of Oregon
Master of Science, Psychology, 2014, University of Oregon
Bachelor of Science, Philosophy-Neuroscience-Psychology, 2010, Washington University in St. Louis

AREAS OF SPECIAL INTEREST:

Self and Identity
Self-Regulation
Social Neuroscience
Experimental Philosophy
Translational Science

PROFESSIONAL EXPERIENCE:

Graduate Teaching Fellow, University of Oregon Department of Psychology, 2012-2018
Graduate Research Fellow, University of Oregon Department of Psychology, 2012-2018
Post-Baccalaureate Research Assistant, Washington University in St. Louis Department of Psychology, 2010-2012

GRANTS, AWARDS, AND HONORS:

Travel Award, Mind and Life Summer Research Institute, 2017

Summer Institute in Social and Personality Psychology, University of Southern California, 2017

Summer Fellowship in Neuroscience and Philosophy, Duke University, 2016

Templeton Sub-Grant in Neuroscience and Philosophy, *Beyond the essential moral self: The importance of social relationships in judgments of first- and third-person identity change*, Duke University, 2016

National Science Foundation Graduate Research Fellowship, *What is ego depletion? A mechanistic exploration of self-regulation and self-affirmation*, University of Oregon, 2013

Poster Prize, Social Affective Neuroscience Society, 2013

Poster Prize, Social Affective Neuroscience Society, 2012

University Honors (*summa cum laude*), Washington University in St. Louis, 2010

Undergraduate Summer Research Grant, Washington University in St. Louis, 2009

NIH Summer Fellowship in Computational Neuroscience, University of Pennsylvania, 2009

PUBLICATIONS:

Livingston, J. L., Kahn, L. E., & Berkman, E. T. (2017). The identity-value model of self- regulation: Integration, extension, and open questions. *Psychological Inquiry, 28*(2-3), 157-164.

Berkman, E. T., Livingston, J. L. & Kahn, L.E. (2017). Finding the "self" in self-regulation: The identity-value model. *Psychological Inquiry, 28*(2-3*), 77-98*.

Berkman, E. T., Hutcherson, C. A., Livingston, J. L., Kahn, L.E., & Inzlicht, M. (2017). Self-control as value-based choice. *Current Directions in Psychological Science, 26*(5)*,* 422-428.

Berkman, E. T., Kahn, L. E., & Livingston J. L. (2016). Valuation as a mechanism of self-control and ego depletion. In E. R. Hirt (Ed.), *Self-Regulation and Ego Control*. New York: Elsevier.

Livingston, J. L., Kahn L. E., & Berkman, E. T. (2015). Motus Moderari: A neuroscience-informed model for self-regulation of emotion and motivation. In G. H. E. Gendolla, M. Tops, & S. Koole (Eds.), *Handbook of Biobehavioral Approaches to Self-regulation* (pp. 189-207). New York: Springer.

ACKNOWLEDGMENTS

There are multiple "selves" that need to be thanked for their support in the making of this dissertation. First, I'd like to thank my dissertation committee, otherwise known as my dream committee: Drs. Elliot Berkman, Jennifer Pfeifer, Rob Chavez, and Nicolae Morar, each of whom studies the self in such novel and diverse ways. A special thanks is due to Elliot who – despite the fact that the only thing in common between the study of self and self-regulation is a single word – found a way to integrate my own research interests with those of the lab, and who taught me that my future "self" need not be a perfect one but simply one who does what she loves.

One of those things I love is incorporating big ideas into my work, and I am particularly grateful for the opportunity to have participated in the Duke Summer Seminars in Neuroscience and Philosophy, a program that funded a large portion of this work through a grant from the Templeton Foundation. The program afforded the opportunity to collaborate with a number of individuals across disciplines, especially Gus Skorburg, who, in an email entitled "A breakthrough?", first proposed that we investigate the moral self.

Neither set of studies would have been feasible without the support of the Social and Affective Neuroscience Lab. Not only has the help with technical aspects of study design, programming, and analysis been invaluable, but the day to day support in in the pursuit of this quirky, abstract topic has truly made "my year." Particular thanks are due to Lauren Kahn, my lab twin, for sharing musings about the self in its many forms throughout the entirety of this journey.

Finally, I am beyond grateful for the love and support of my family and friends who trusted me in my decision to move to a small, Oregonian town in pursuit of adventure and knowledge. The cast of characters and scenes that have extended the boundaries of myself will not be forgotten – they will always be a part of me.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

Table Page

CHAPTER I

INTRODUCTON

Definitions of self and identity within psychology and neuroscience are "extremely wide- ranging and [lack] uniformity" (VandenBos, 2007). Researchers in the field have approached the study of self from many directions, the result of which has been an outpouring of self-hyphenated terms: self-regulation, self-esteem, self-knowledge, self-concept, self-perception, and more (Katzko, 2003; Klein, 2012a). Despite their similar root structure, the degree to which these words reference the same "self" or aspect of "self" remains unclear, leading some to claim that, based on philosophical grounds, the word has even lost its meaning (Wittgenstein, 1953; Bergner, 2017). At the broader level, the study of self lacks clarity and coherence – a broader organizational structure that gives the study of self purpose – a structure and purpose to which the current dissertation aims to contribute.

Ironically, even though approaches to the psychological study of self have lacked coherence, most approaches have been limited in similar ways. First, most research examining the self has approached the self as a single, unified construct at the expense of either ignoring or failing to integrate approaches that view the self as multidimensional (e.g., McConnell, 2011). This idea is implicit in many of the self-hyphenated terms listed above. Second, most research examining the self has studied the self at one moment in time (i.e., "synchronic" self) at the expense of failing to examine the continuity of self over time (i.e., "diachronic" self) (Northoff, 2017). As a result, much of the seemingly

wide-ranging psychological literature on self has been limited to a relatively narrow subset of information.

The same limitations have applied in the neuroscience literature, perhaps even more so. The first investigations of the neuroscience of self utilized a task that prompts participants to rate the extent to which different personality trait words described themselves (Craik et al., 1999; Johnson, Baxter, Wilder, Pipe, Heiserman, & Prigatano, 2002; Kelley, Macrae, Wyland, Caglar, Inati, & Heatherton, 2002). Participants completing this task are presented with single prompts (e.g., "Self" or "Friend") that indicate the target of reference, as well as single trait words (e.g., "Polite" or "Talkative"), and are asked to indicate whether or not the trait word describes the target. In the same way that the psychological study of self has been limited, this task is limited in that the traits presented are implied to describe a single, de-contextualized self, neglecting other dimensions, such as time and social roles, that could be critical for constructing a comprehensive neural representation of self. The trait-based tasks were presumably designed as first attempts to address the topic and were derived from the somewhat limited, strictly cognitive perspective at the time (e.g., tasks adapted from the memory literature). Many studies have since examined the neural activity underlying self-relevant cognition in a number of ways (Denny, Kober, Wagner, & Ochsner, 2012; Wagner, Haxby, & Heatherton, 2012), but the gold standard in the field, the task that is used to reliably elicit self-referential activity, is the same, trait-focused task used in these first studies (e.g., Moore, 2015).

Different frameworks have been proposed as an attempt to organize and advance the comprehensive study of self. Many psychologists have pointed to the Jamesian

knowing, subjective "I" versus the known, objective "me" distinction as a helpful framework for organizing the functional relationships between the "selves" in the self-hyphenated terms. Even so, most traditional studies of self and identity in the field have investigated the "me," (i.e., the *content* of the self) without considering how it relates to the "I," despite suggestions that the systems likely are not as separate as they may seem (Northoff, 2007; Ochsner & Gross, 2007). However, clarifying this relationship is difficult without first acknowledging the full scope of self-relevant content (e.g., actor, agent, and author) that allows the I to conceive of the Me (McAdams, 2013).

The full scope of self-relevant content can be captured within a framework that identifies the various levels at which we can know and describe a person (McAdams, 1995). The first level at which we can know someone consists of the familiar trait approach that attempts to characterize an individual's broad, decontextualized disposition. For example, at this level, a person might be described as "polite" or "talkative," but these traits are limited in that they are both *nonconditional* – that is, they do not take the role of context into account, as well as *comparative* – that is, they are designed to tell us how an individual compares to others on different trait dimensions, but they do not tell us much about the content that uniquely identifies a person over time. This is also the level at which most neuroscience studies have investigated the self.

The second level under this approach identifies an individual's "personal concerns," such as their goals, motivations, and values that are more contextualized to particular times and roles. For example, at this level, we might learn that an individual is talkative because they are uncomfortable with silence but that they are striving to listen

more and talk less. This arguably provides observers with more information that uniquely characterizes this person compared to the first level.

The final level in this model identifies the narrative self that individuates a person, generates meaning and purpose for that person, and integrates elements of an individual's life experience across both content and time. Description at this level consists of detailing the key scenes, characters, and themes that characterize an individual's life in important domains, such as this talkative individual's autobiographical experiences of engaging in meaningful and mundane conversations with others. Importantly, the framework claims that we cannot know a person, know their self or their identity, without considering information from each of these levels. Even so, consideration for the information involved at these final two levels has been relatively lacking from the literature on self-relevant processing.

Clarifying information at these final two levels is important for a number of reasons. First, from a functional perspective, clarifying the nature of the self is critical for understanding how identity can be used as an effective motivational tool to help individuals pursue their goals (Berkman, Livingston, & Kahn, 2017). Questions surrounding the self and its role in directing intentional, goal-oriented behavior are perennial in both philosophy and psychology. In contrast to the Platonic notion of self-control in which a metaphorical charioteer is challenged with the task of steering one rational and one impulsive horse in the same direction, many contemporary philosophers point towards the role that the value-laden self can play in guiding successful, controlled behaviors. For example, it has been argued that intentional behavior is feasible strictly

because individuals use their own values to guide their decisions (Bechtel, 2008) and that willful behavior requires the ability to engage the desires of the self (Frankfurt, 1988).

Such philosophical perspectives are in accordance with a new, value-based model of self-control that aims to improve upon existing dual-process models of self-control (Berkman, Hutcherson, Livingston, Kahn, & Inzlicht, 2017). To date, most studies investigating self-control have operated under a dual-systems approach that defines self-control as a competition between hot, impulsive desires and cold, effortful driven control. Although dual systems models fit well with the phenomenological experience of engaging in self-control, they cannot account for a number of behavioral findings in the self-control literature, such as ego depletion (Berkman, Kahn, & Livingston, 2016); nor is there strong evidence for a negative relationship between control-based regions and reward-based regions of the brain to support this model (Kelley, Wagner, & Heatherton, 2015). In contrast, the value-based model of self-control draws upon evidence from the neuroeconomics literature to propose that self-control, or the act of choosing amongst any number of mutually exclusive options for the purposes of engaging in goal-oriented behavior, can be thought of as the output of a value-based calculation amongst competing options. In this model, valuation is a core mechanism of self-control that can parsimoniously account for a wide variety of psychological phenomena (e.g., the endowment effect, temporal discounting, etc.) and that is extensively supported by the existing neuroscience literature (Hutcherson, Bushong & Rangel, 2015; Kable & Glimcher, 2007).

The implication of the model is that, because higher valued inputs determine behavioral output, interventions aimed at motivating successful goal pursuit can increase

the relative weight of relevant value inputs. There are a number of potential avenues to doing so (e.g., monetary value, social support), but the psychological literature suggests that perhaps the most salient route to high value is identity. Not only is identity both chronically accessible as well as stable across time (Markus & Kunda, 1986), but it also tends to be positive in nature (Rosenberg, 1979; Steele, 1988), at least in healthy individuals. Indeed, evidence from within psychology supports the idea that identity can facilitate different forms of successful self-regulation (Schmeichel & Vohs, 2009; Ersner-Hershfield, Wimmer, & Knutson, 2009; Higgins, Roney, Crowe, & Charles, 1994). Moreover, recent neuroimaging studies note strong overlap between neural activity associated with thinking about identity and value (Kim & Johnson, 2015, Northoff & Hayes, 2011), activity which is also often predictive of successful goal-pursuit (Berkman & Falk, 2013). In fact, a meta-analysis and conjunction using a database (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011) of neuroimaging studies on self and value reveals a large cluster of overlapping activation in the ventromedial prefrontal cortex (vmPFC), suggesting that the two constructs may be intimately related, if not one in the same. As a result, the identity-value model of self-regulation extends the value-based model of self-control to propose that identity serves as particularly promising target of intervention (Berkman, Livingston, & Kahn, 2017).

The identity-value model has been well-received as an important contribution to the literature (e.g., Hackel & Zaki, 2017; Lempert & Kable, 2017); however, critical questions about the nature of identity remain unexplored (Livingston, Kahn, & Berkman, 2017). Perhaps most critically, if identity is to serve as a target of intervention, it is unclear how identity can remain at once stable and flexible; that is, it is unclear how

identity can serve as a potent and accessible source of value while remaining amenable to intervention. The model's seemingly contradictory demands are nearly impossible to satisfy when assuming the dominant, unitary account of self in the literature. A more comprehensive account that considers the self across different contexts and across time, however, might be better able to simultaneously satisfy the identity-value model's claims.

Not only can understanding information about the self across context and time inform functionalist approaches, but investigating the self at these final two levels can help to advance ontological questions about the nature of the self. Many of these key questions center around how the self is unified across both content and time (Klein, 2012b), and philosophers have a number of frameworks for addressing these types of questions. A comprehensive review of these frameworks lies outside of the scope of this dissertation, but two distinctions are particularly helpful for studying the self across context and time. First, with regards to context, William James (1890) is careful to distinguish between the social selves that constitute roles for the self and the spiritual self that constitutes the perception of a unified self. Such a distinction proves particularly valuable when considering the relationship between empirical approaches that examine an individual's multiple selves (see Chapter 2) and more traditional approaches that assume a unified self. Second, with regards to time, recent work is a careful to distinguish between the narrative self, the collection of the identities that one possesses and personal identity, the core subject that possesses all of those identities (Peacocke, 2014). Such a distinction proves particularly relevant when considering the relationships between empirical approaches examining personal identity (see Chapter 3) and broader narrative-based approaches to studying the self.

Integrative approaches that attempt to unify seemingly separate conceptions of self to meet broader functionalist and ontological needs have the potential to advance the study of self. Although the advantages of an integrative approach might appear obvious to some, calls for integrated approaches across the humanities and sciences are often met with either "indifference or hostility" (Slingerland & Collard, 2011). However, the advantages of such integrative approaches are recently becoming evident. Despite the limitations of the original studies that investigated the neuroscience of self (discussed earlier in this chapter), these studies were, in many ways, integrative. For example, one of these studies (Kelley et al., 2002) applied neuroscience methods to address a question at the nexus of cognitive science and social psychology: namely, is memory for the self special? Although the types of claims that can be drawn from the trait-based methods used in this study are, in many ways, limited, the findings from the study have recently provided the groundwork for a number of important discoveries about the self, including the discovery that neural activity about the self can be predictive of real-world behavior (Berkman & Falk, 2013) and that neural activity about the self is highly overlapping with value-based activity (Berkman, Livingston, & Kahn, 2017). Extending the study of self to include more role-based and time-based perspectives has the potential to push this groundwork even further.

The current dissertation contributes towards an integrative approach by extending the study of self-relevant processing to consider the role of context and time. Chapter 2 explores elements of self-complexity and reports the results of an empirical investigation on multiple selves. Chapter 3 discusses elements of self-continuity and reports the results of an empirical investigation on personal identity and the moral self. Finally, Chapter 4

discusses and integrates the results from Chapters 2 and 3 to consider how these studies contribute to the study of the self across context and time.

This suite of papers extends traditional approaches towards the study of self in a number of important ways. Notably, by transitioning away from trait-centered approaches and toward goal- and narrative-based approaches, the dissertation steers away from the traditional synchronic perspective on self and towards a diachronic perspective on self that considers contextual roles and changes over time. Moreover, the dissertation's methodologies pull from a variety of disciplines, both philosophical and translational, to achieve the projects' aims. The result is a contribution towards an integrative study of self, both in method as well as in result.

CHAPTER II

MULTIPLE SELVES

One Self or Many?

In his self-explorative tome *In Search of Lost Time*, Proust declares "I have developed the habit of becoming myself a different person [...]: a jealous, an indifferent, a voluptuous, a melancholy, a frenzied." Proust's different persons are contextually driven elements that are stored in memory as remnants of selves (Landy, 2001). In other words, this Proustian self which so famously characterizes the human experience is one that also maintains a many-tiered structure of sub-selves (Kemp, 2005).

Unlike Proust, most psychological research exams the single, unified self at the expense of failing to examine its multiple, context-dependent selves (e.g., McConnell, 2011). Although this preference for studying the unity of self over its multiple aspects is rarely made explicit (e.g., Baumeister, 1998), the tendency is implicit within many sub-domains of psychology. For example, research on self-protective mechanisms assumes a unified self when examining the effects of exercises that affirm core, unifying values on self-integrity in the face of threats (Steele, 1988). Self-esteem research assumes a unified self when asserting that there is a general, unified, positively affected feeling one can have towards the self (Rosenburg, 1979). And many lines of research on accuracy (e.g., self-knowledge) assume a unified self in asserting that there is one, true self that we either do or do not come to know (Vazire & Carlson, 2010). Much of the seemingly wide-ranging psychological literature on self has been limited to a relatively narrow definition.

The multiple self-aspects framework suggests that, contrary to the traditional assumptions of a singular self in psychology, the self is composed of multiple, context-dependent self-aspects (McConnell, 2011). Despite its non-dominant status, the perspective that the self can be broken into sub-components is not new. Multiple-self perspectives date back at least as far as William James who asserted that we have "many social selves" (1890), and since James, numerous dichotomies and trichotomies dividing the self into different aspects have emerged. The self can be divided into the past, present, and future self (D'Argembou, Stawarczyk, Majerus, Collette, Van der Linden, & Salmon, 2010), tracking its course through time, and into the actual, ought, and ideal self, tracking its aspirational statuses (Higgins, 1987). It can be divided into the independent and interdependent self-elements, tracking relationships with other individuals (Markus & Kitayama, 1991), as well as into different social roles (Roberts & Donahue, 1994). However, these existing frameworks divide the self into separate, pre-specified dimensions of self, lacking any form of integration across the dimensions.

The developmental literature has also long recognized that the self is fragmented. This perspective is inherent in the idea that the self is constructed over time, such that self-perception becomes increasingly more complex as a child assumes new social roles and develops new relationships (Harter, 2012). The process of self-construction is particularly salient during adolescence when younger adults are still exploring many aspects of their self and identity, including their goals, motivations, and responsibilities (Becht et al., 2016; Meeus, Iedema, Helsen, & Vollebergh, 1999). A key aspect of development, then, is the acquisition of differentiated, domain-specific self-concepts (Pfeifer & Berkman, in press), but, to date, the study of the self as just that – a collection

of different self-concepts - has been relatively lacking from the social psychological literature.

The Multiple Self-Aspects Framework

The multiple self-aspects framework directly addresses this lack of consideration for the differentiated self. Importantly, rather than constraining the boundaries of self to any pre-specified divisions, the multiple self-aspects framework allows elements of self (i.e., self-aspects) to vary idiosyncratically across individuals. The framework is inspired by evidence suggesting that when individuals are asked to sort trait words into categories that represent meaningful aspects of their lives, they typically identify four to five key self-aspects comprised of various social roles (e.g., student self), identities (e.g., sorority self), relationships (e.g., girlfriend self), goals (e.g., ideal self), or affective states (e.g., stressed self) that organize their experiences and help to direct their actions (McConnell & Strain, 2007). Notably, these self-aspects are idiosyncratic in that they are uniquely identified by each individual, a development that distinguishes this approach from the other lines of work that only examine particular, pre-specified dimensions of the multifaceted self.

The multiple self-aspects framework has a number of other unique features. In particular, not all trait words are used for each participant, and some of the same trait words are used to describe different self-aspects, yielding a highly individualized self-knowledge structure that overlap by different amounts across individuals. Within individuals, self-aspects are accessible to varying degrees at any one given time (Brown, Bailey, Stoll, & McConnell, 2016), accessibility is malleable and context-dependent such that different self-aspects are accessible to varying degrees in different situations, and the

relative accessibility of a node in a given moment impacts the self's overall affect (i.e., experienced positivity of the self) (McConnell, Rydell, & Brown, 2009). What emerges is an integrated collection of self-aspects that are relatively stable but whose relationships are flexible across time.

To date, this organizational structure has been measured using a single metric of self-complexity. This metric takes into account both the total number of self-aspects listed by an individual, as well as the degree to which an individual's self-aspects share trait attributes with one another, using an H-statistic (Scott, 1969): $H = \log_2 n - (\sum n_i \log_2 n_i) / n$, in which n represents the total number of traits presented to the participant and n_i represents the number of traits that fall under each self-aspect grouping (i) identified by the participant (McConnell et al., 2005). An individual who has a higher number of self-aspects that share less attributes with one another is said to have high self-complexity, and an individual who has a lower number of self-aspects that share more trait attributes with one another is said to have low self-complexity. Research in this area suggests that self-complexity is a measurable phenomenon within an individual that varies across individuals in systematic ways.

Self-Complexity and Well-Being

Measuring self-complexity is important not just for validating ideas within psychology, but also for predicting real-world outcomes. Meta-analysis reveals that individuals with higher self-complexity generally experience worse physical and psychological health outcomes than individuals with lower self-complexity (Rafaeli-Mor & Steinberg, 2002), although it is not entirely clear why lower self-complexity seems to confer advantages, on average. Some work has suggested that perceived self-aspect

control moderates the relationship such that lower self-complexity is associated with better well-being insofar as it helps individuals feel more in control of their self-aspects (McConnell et al., 2005). This perspective is consistent with clinical approaches that diagnose individuals with dissociative identity disorder to the extent that they lack consensus and integration between their various selves and report feeling out of control (Vanderlinden, Van Dyck, Vandereycken, Vertommen, & Jan Verkes, 1993).

Although lower self-complexity is associated with well-being on average, its advantages may be context-dependent. Whereas lower self-complexity is associated with well-being for individuals who already have positive influences in their life, such as a social support network and socially desirable personality traits, it is also associated with poorer well-being for individuals who have experienced negative life events (McConnell, Strain, Brown, & Rydell, 2009). Such findings are in line with the buffering hypothesis, which states that high self-complexity can be helpful for buffering the self from threats (Linville, 1985). Indeed, evidence suggests that multifaceted selves can buffer mood in response to negative self-feedback (Linville, 1985; Dixon & Baumeister, 1991), can reduce distress in response to a breakup (Smith & Cohen, 1993), and can help individuals cope in the face of trauma (Nijenhuis & van der Hart, 2011; DePrince & Freyd, 2014). Individuals greater in self-complexity, it seems, are better able to resist threats to one self because their selves, as a whole, are not integrated.

Malleability of the Self

To date, then, self-complexity has been examined as a stable individual difference measure that predicts the extent to which individuals experience certain health outcomes (Rafaeli-Mor & Steinberg, 2002). Given the relationship between self-complexity and

measures of well-being, however, intervention work may want to consider targeting self-complexity in order to facilitate well-being. Of course, the relationship may be bi-directional, such that changes to well-being may affect self-complexity. However, a recent model of self-regulation hypothesizes that targeting identity is an effective way to achieve certain forms of well-being (Berkman, Livingston, & Kahn, 2017).

The main advantage of targeting identity is that it is both stable, providing a strong source of value, and flexible, or amenable to intervention. On the surface, these characteristics seem incompatible with one another (Molden, Hall, Hui, & Scholar, 2017), but only if one assumes the self to be a unified entity. Fortunately, the multiple self-aspects framework provides two potential stable and flexible sources of identity which may be differentially amenable to intervention: the self-aspects themselves and the relationships between them.

There are techniques psychologists may be able to use to try to introduce new self-aspects or to manipulate existing self-aspects and/or the way they relate to one another. For example, cognitive dissonance hypothesizes that motivating someone to engage in a new behavior may lead them to justify that new behavior by incorporating it into their identity (Festinger, 1957). Similarly, certain persuasive techniques, such as the availability heuristic, can be utilized to draw attention to certain behaviors for similar effects (Schwarz, Bless, Strack, Klumpp, Rittenauer-Schatka, & Simons, 1991). Contemporary philosophers have even theorized that elements of identity can be "summoned" in a self-fulfilling prophecy, such that individuals, in a sense, simply become what they are called by others (Alfano, 2015). Although the idea is intriguing, psychologists have also long recognized the role of autonomy in behavioral change (Deci

& Ryan), challenging the notion that new identities can be feasibly introduced or nudged from outside sources.

Rather than introducing new self-aspects, self-complexity may be more easily modified by targeting the relationships between existing self-aspects. A limited subset of existing strategies from the self-regulation literature have demonstrated that modifying the way that individuals relates to themselves can promote behavior change. For example, self-distancing, which requires individuals to maintain a removed perspective from the self (e.g., talking to the self in the third person) can effectively dampen emotional reactivity (Kross & Ayduk, 2011). Self-affirmation, which instructs individuals to think about a top-ranked value to the self, can encourage individuals to engage with health-promoting information (Klein & Harris, 2009). And self-compassion, which encourages individuals to talk to themselves with care, as they would another person, has been shown to promote resilience to negative life events (Leary, Tate, Adams, Batts, & Hancock, 2007). Despite the effectiveness of these interventions, they seem to act upon the self as a whole. Indeed, one of the mechanisms hypothesized to unify all of these self-oriented regulation strategies is cognitive abstraction, a form of psychological distance as achieved through high-level construal (Sklar & Fujita, 2017), which is a well-understood cognitive technique that prompts people to focus their attention on the global, primary features of an object or event in order to facilitate self-control (Fujita, Trope, Liberman, & Levin-Sagi, 2006). As a result, it's unclear how these manipulations might impact self-complexity. An intervention targeted at particular elements of the self may be more effective.

Although psychologists have traditionally focused on the self, as a whole, in researching self-regulatory techniques, various therapeutic techniques have focused on targeting different components of the multi-dimensional self. Third wave therapies, in particular, have focused on self-integration, not only across therapies, but also across the aspects of a given individual. For example, integrative psychotherapy focuses on integrating elements of personality into a coherent whole (Erskine & Trautmann, 1993). This emphasis on integration is in line with certain philosophical perspectives that identify integration as a key, desirable feature of personhood (Frankfurt, 1971) Other third wave therapies nod to the multidimensional nature of the self by acknowledging and addressing the particular needs of an individual in a given context (Hayes, 2004).

Rather than changing self-aspects themselves, then, researchers may want to focus on developing tools that can target the relationships between existing self-aspects. Previous work has successfully targeted these relationships by using priming techniques to shift the relative accessibility of different self-aspects (McConnell, Rydell, & Brown, 2009). However, it's unclear how long-lasting these priming effects are. Rather than targeting accessibility, interventions may want to focus on targeting the trait level attributes associated with each of the self-aspects. Although traits are relatively stable (Caspi, Roberts, & Shiner, 2005), they are also known to shift under different contexts (e.g., Roberts, Luo, Briley, Chow, Su, & Hill, 2017). Thus, changing the relative associations between self-aspects, and their traits, across different contexts may be a viable method for encouraging individuals to think about themselves in ways that are beneficial to their well-being.

The Case for Networks

A prerequisite for any meaningful scientific work focusing on the structure of the self is a nuanced measure that is capable of detecting shifts in the relationships between self-aspects. Whereas the h-statistic, the traditional measure of self-complexity, calculates scores based on binary (i.e., "yes" or "no") self-categorizations, it is less sensitive to the changes of degree that would be expected from within-subject changes over time. For example, it is unlikely that a person who is not confident (i.e., a "no") in in their athletic ability would suddenly rate their athletic self-aspect as "confident" (i.e., a "yes") after a manipulation – the type of change required for the h-statistic to detect any differences; however, it's possible someone might rate their athletic self-aspect as slightly more confident (e.g., an increase from a "2" to a "3" on a 7-point scale) after considering the ways in which their athletic self-aspect integrates with other aspects of their life.

The H-statistic is also limited in that it cannot measure any potentially nuanced relationships between individual self-aspects. Although previous literature has indicated that lower self-complexity is, on average, associated with better well-being, subtle variations in how self-aspects relate to each other are likely meaningful to outcomes including, but not limited to, well-being. For example, because less overlapping self-aspects can be helpful in the face of threat, it might be optimal for positive self-aspects to share overlapping qualities (low self-complexity) but for negative self-aspects to share less overlap (high self-complexity). Indeed, a parallel line of reasoning from the self-affirmation literature in which the self, as a system, is motivated to protect its integrity in the face of threats would support this hypothesis (Sherman & Cohen, 2006). Bolstering (or "affirming") different aspects of the self that are only weakly related to the threatened aspect of the self can protect against the deleterious effects of self-threat. An intervention

that either encourages integration or multidimensionality, then, may only be interested in measuring changes in overlap between particular traits. A more nuanced method for capturing these subtler relationships is needed.

Tools derived from the branch of mathematics known as graph theory can be used to create visual maps called 'networks' that summarize people's reports about their self-aspects and uses information about the relationships between those self-aspects to generate related but separate metrics about those networks. Networks are made up of a collection of nodes (self-aspects) and edges (relationships between self-aspects). Edges can be defined in a variety of ways, but in many cases, they are defined by the correlation between two nodes. Tools from graph theory (Bondy & Murty, 1976; West, 2001) can be used to construct network maps and to calculate overall network-level metrics, such as size and clustering, from these networks. Size can be measured in a variety of ways, such as by simply counting the number of total nodes (self-aspects) in a network, and clustering indicates the extent to which nodes (self-aspects) cluster together into groups. These types of measures are helpful for understanding information about how all self-aspects within a network relate to one another.

Tools from graph theory can also be used to calculate node-level (self-aspect level) metrics, such as strength, betweenness, and closeness. Strength indicates information about the number of (weighted) connections for a particular self-aspect, betweenness indicates the extent to which a given self-aspect is located in the middle of connections between other self-aspects, and closeness provides information about the shortest (weighted) distances between a given self-aspect and all other self-aspects in the network. All of these measures can provide meaningful information about how important

a particular self-aspect is to a network (Costantini et al., 2014). Therefore, network approaches would likely be useful for both characterizing relationships between specific self-aspects, as well as for measuring their change over time.

Insights from network-based approaches have been able to push psychological theory forward in a variety of domains. Applying network-based approaches to the study of personality, for example, has generated data that challenged long-held assumptions within the personality literature, which traditionally relies upon latent variable approaches (Costantini et al., 2014). And it's been suggested that network approach would be similarly appropriate for examining relationships between identities within a person (Ramarajan, 2014). Applications of network analysis within neuroimaging has also advanced thinking on a number of topics, including the ability to identify collections of brain regions involved in certain neuropsychiatric disorders for particular individuals (Bullmore & Sporns, 2009). Also, the correlation matrices upon which these network analyses rely serve as the basis for a suite of other brain decoding techniques, many of which have revealed important insights about the structure of the self (e.g., Chavez, Wagner, & Heatherton, 2017).

The Present Study

The current set of studies extends the work on multiples selves in a number of important ways. Study 1 serves as a conceptual replication of the work by McConnell et al. (2005) on the relationship between self-complexity and well-being in an online, non-student population, utilizing a continuous rating scale rather than a card sorting task. Study 2 extends the effect by developing and applying more nuanced graph theoretical techniques to build self-networks and by using them to measure self-complexity and its

relationship to well-being. Finally, Study 3 examines whether and how self-aspect structures can be both stable across time and also amenable to manipulation. This suite of studies aims not only to develop a novel method for investigating self-aspect structure, but to also contribute toward a broader literature thinking about the multiplicity of selves and the tools that can be used to measure the self in its many forms.

Study 1: Replication

Method

Open science. The design, hypotheses, and analysis plan for this study were pre-registered at the Open Science Framework (https://osf.io/t3hvj/).

Participants. Participants (18 years of age or older, native English speaker, U.S. resident) were recruited from the Amazon Mechanical Turk population using TurkPrime's data acquisition platform (Litman, Robinson, & Abberbock, 2017) to participate in a 30-minute online study inviting participants to make ratings about themselves. Sample size was determined through a power analysis using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) showing that a sample size of 506 (253 per condition) would be required to detect a small effect (d = 0.25) in an independent samples t-test with 80% power. This is conservative because the primary analysis focuses on a mixed factorial ANOVA analysis, so the actual analysis might have more power than a simple t-test. A 30% drop out rate in Study 3 as well as 10% data loss due to attention checks was estimated to determine the number of observations to gather. This sample size well-exceeds the size of the sample (N = 127) of the study that my study aims to replicate (McConnell et al., 2005). Participants were excluded if they did not complete the survey

or failed either of two other attention checks included in the survey. The final sample for analysis included 789 participants (403 female, *Mage* = 38.29, *SD* = 12.06).

Design. In this survey, participants were asked to generate a list of their self-aspects and to make trait ratings for each, with no pre-specified conditions.

Procedures. After consenting, participants generated a list of their self-aspects using instructions adapted from Showers (1992). A self-aspect was defined for participants as something that "identifies an important aspect of yourself or your life," and were told to consider themselves across "different situations, relationships, roles, emotions, goals, and time periods." Participants were presented with an example (Harry Potter) for reference but were encouraged to include any self-aspects that were meaningful to them. Participants were instructed to list as many or as few self-aspects as desired, were told that most people tend to list four to five self-aspects, and were required to list at least two (but no more than eight). Moreover, participants were notified that after generating their self-aspects, they would rate each of their self-aspects on both positive and negative trait words.

Then, participants spent at least two minutes listing their self-aspects and were unable to proceed to the next section until two minutes had passed. In the next section, each individual's self-aspects were then piped into a question in which participants were asked "to what extent does the following word describe your _____ self-aspect?" (1 = not at all to 7 = very much) for 40 different randomly presented traits words (20 positive and 20 negative) previously established as common self-descriptors (Showers, 1992) (Appendix A). The questions were presented in blocks so that participants answered all 40 questions for one self-aspect before moving on to the next self-aspect, with order of

self-aspect blocks randomized (Figure 1). At the end of each block, participants were asked three more questions about overall centrality, positivity, and negativity for each self-aspect.

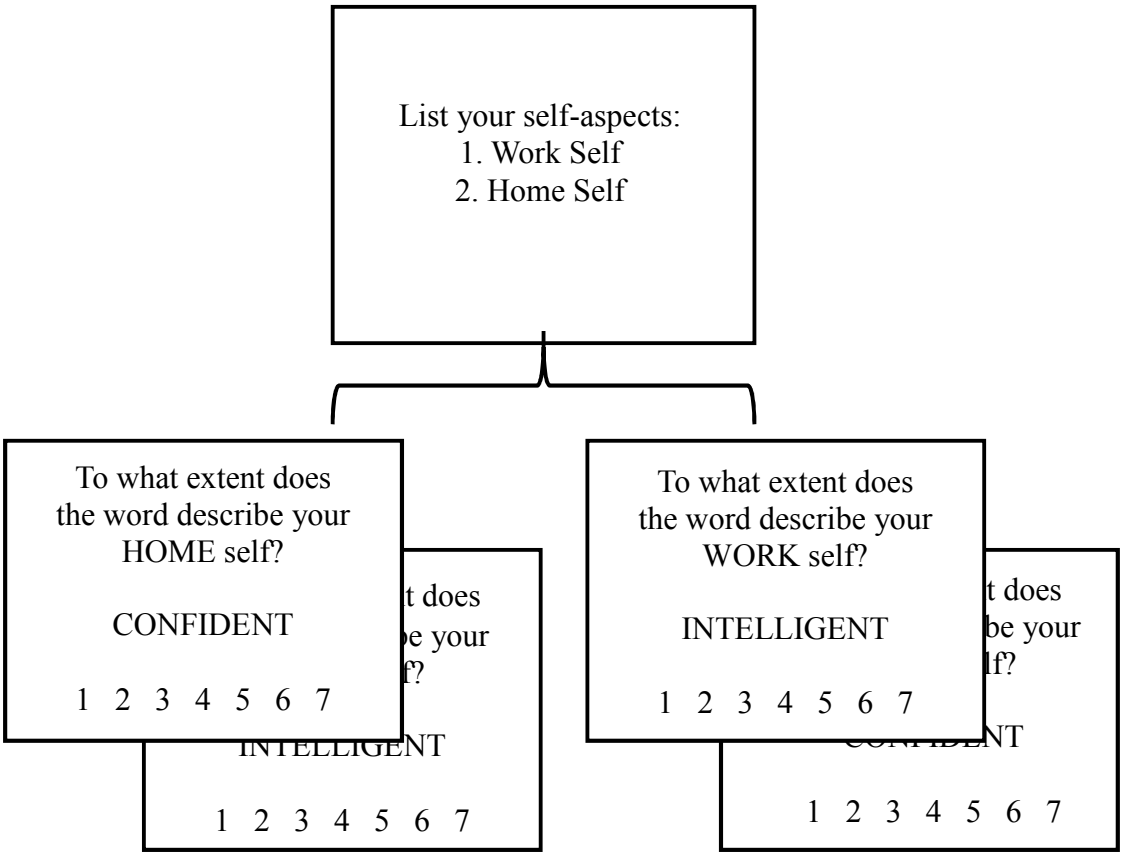

Figure 1. Task procedure for generating and rating self-aspects. Participants were asked to list their self-aspects (up to eight) and then were asked to rate each self-aspect on each of the traits.

Additionally, participants were asked to complete a variety of individual difference measures pertaining to physical well-being, including the Cohen-Hoberman Inventory of Physical Symptoms (Cohen & Hoberman, 1983) and the Perceived Stress Scale (Levine & Perkins, 1980) Cohen, Kamarck, & Mermelstein, 1983), measures pertaining to psychological well-being, including the Rosenberg Self-Esteem Scale

(Rosenberg, 1965) and the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), as well the Self-Concept Clarity Questionnaire (Campbell, Trapnell, Heine, Katz, Laveallee, & Lehman, 1996), followed by free response questions asking if they would have liked to add, remove, or change any of their listed self-aspects, and a demographic questionnaire. Finally, participants were presented with a debriefing form describing the study and a code to submit to MTurk in exchange for payment. Given that some participants listed more self-aspects than others and took longer to complete the survey, any participants who took longer than the estimated time were provided additional bonus payment (paid at an established hourly rate).

Analysis plan. The primary analysis for this study sought to replicate results from McConnell et al., 2005. This study is exploratory in the sense that the *a priori* single best way to analyze the data was unknown. To enable experimentation with different analytic approaches without overfitting to the data, all of the analyses were first performed and refined on a "training" subset of the data ($N = 491$). Once the analyses were established and finalized on the training sample, they were applied and verified on a held out "test" sample of data ($N = 298$). Data from the entire sample were randomly assigned to either the training or test sample so as to avoid any systematic differences between the samples, with assignment stratified by the time 2 condition assignments (see Study 3) such that both conditions were represented equally in both the training and test sample.

Conceptual replication was tested using the same metrics as reported in the original study (McConnell, 2005), including the calculation of the h-index (Scott, 1969), which weights both the number of self-aspects listed and the degree of overlap to measure self-complexity, as well as the correlation between the h-statistic and a number

of well-being measures. However, given that I decided to use a continuous scale rather than a categorical trait judgment (i.e., yes or no) to measure the relation of each trait to each self-aspect, trait judgments were transformed to a categorical scale for the calculation of the h-statistic. I used a variety of methods for doing so (e.g., rating threshold) in the training sample and applied only one to the test sample. The h-index that most closely replicated McConnell et al. (2005) was selected, while also considering the shape of the distribution and variance for each option. The selected transformation was then applied to the test sample, and the resulting mean self-complexity (h-index) scores, as well as their correlations with all of the well-being measures were calculated. To examine the role of affect in self-complexity, the correlation between mean overall self-aspect positivity and the well-being measures was also examined.

Results

I hypothesized that the mean number of self-aspects reported in McConnell (2011) and the mean self-complexity score reported in McConnell et al. (2005) would replicate, as would the correlations between the mean self-complexity score and the well-being measures. On average, across the training and test sample combined, people reported having approximately 5 self-aspects ($M = 4.92$, $SD = 1.69$), replicating results previous results.

Iterative testing was run in the training data to evaluate different methods for binarizing the continuous scale trait ratings. This testing revealed that categorizing trait responses of 6 or 7 as "yes" and of 5 or lower as "no" fit the original results most closely (ratings were made on a 1-to-7 scale where 1 indicated "not at all" and 7 indicated "very much"). A one-sample t-test on the training data revealed that a threshold of 6 was the

only threshold that yielded an h-statistic (M = 2.34, SD = 0.80) that did not significantly differ from the h-statistic (M = 2.33, SD = 0.84) reported in McConnell et al. (2005) ($t(490) = 0.54$, $p = .59$). Moreover, the correlation values associated with this threshold in the training data were most similar in direction and magnitude to those reported in McConnell et al. (2005) (Table 1). Table 1 reveals that greater self-complexity in the training data (using an H-threshold of 6) was negatively correlated with positive markers of well-being, such as self-esteem (M = 30.94, SD = 6.84), and was positively correlated with negative markers of well-being, such as depression (M = 30.34, SD = 11.14) and perceived stress (M = 24.96, SD = 8.45). Notably, a threshold of 5 in the training data (not included in Table 1) yielded an h-statistic that actually correlated more strongly (in magnitude) with the well-being measures than the correlations reported in McConnell et al., 2005. However, given that the goal was to replicate McConnell et al (2005) as closely as possible and that a threshold of 6 was the only threshold to yield a similar h-statistic to that reported in McConnell et. al. (2005), a thresholding of 6 was selected for use in the test sample.

The same analyses were run in the test sample using a threshold of 6 to calculate the h-statistic (i.e., self-complexity). A one-sample t-test comparing the mean h-statistic in the test sample (M = 2.32, SD = 0.82) to the mean h-statistic reported by McConnell et al. (2005) revealed no statistically significant difference ($t(295) = -0.31$, $p = 0.76$). Moreover, the correlation values associated with the h-index in test sample were, for the most part, similar in direction and magnitude as those reported in McConnell et al. (2005) (Table 1). Notably, greater self-aspect positivity was positively correlated with more desirable aspects of well-being such as self-esteem (M = 30.09, SD = 7.29) and

negatively correlated with less desirable elements of well-being such as depression (M = 31.56, SD = 11.92), stress (M = 25.81, SD = 8.46), and physical symptoms (M = 52.26, SD = 19.88), as was also reported in McConnell et al., (2005) and the training data.

Table 1
Intercorrelations among self-complexity (H), self-aspect positivity (Pos.), and well-being measures from McConnell et al. (2005), Study 1 training data, and Study 1 test data

	McConnell et al. (2005)		Study 1 (Train)		Study 1 (Test)	
	H	Pos.	*H*	Pos.	*H*	Pos.
Self-complexity (*H*)	-		-		-	
Self-aspect positivity	-0.28**	-	-.12**	-	-.10	-
Self-esteem	-.16	.50**	-.12**	.49**	-.07	.56**
Depression	.29**	-.43**	0.23**	-.39**	0.17**	-.49**
Physical Symptoms	.23*	-.24**	.22**	-.25**	0.16**	-.26**
Perceived Stress	0.1	-.19**	.16**	-.42**	0.14*	-.49**
Self-aspect negativity	-	-	.26**	-.82**	0.25**	-.86**
Self-aspect centrality	-	-	-0.07	.57**	-0.06	.65**
Self-concept clarity	-	-	-.18**	.36**	-.13*	.49**

Note. *p < .05,
**p<.01.

Discussion

I hypothesized that the findings of McConnell et al. (2005) would (conceptually) replicate such that greater self-complexity would correlate negatively with desirable well-being measures (e.g., self-esteem) and positively with undesirable well-being measures (e.g., depression and stress). At a basic level, the mean number of self-aspects reported by participants in this online sample replicated the mean number reported in previous work (McConnell, 2011). Moreover, as expected, a calculation of self-complexity using established methods (h-index) replicated self-complexity values reported in McConnell et al., 2005. Not only did the self-complexity value replicate, but so did its correlation with

well-being: greater self-complexity was negatively correlated with desirable well-being measures in the training sample (e.g., self-esteem) and positively correlated with undesirable well-being measures (e.g., depression, stress, and physical symptoms) in both the training and the test sample. The correlations within the training sample were similar in both strength and magnitude to those reported in McConnell et al., 2005, whereas those within the test sample were somewhat weaker. For example, although self-complexity was significantly negatively correlated with self-aspect positivity in McConnell et al. (2005) and in the training sample, the negative correlation was not significant in the test sample. This lack of consensus between the training and test sample may partially be a testament to over-fitting in the training sample or to the smaller size of the test sample. It's also possible that using another threshold to calculate self-complexity (e.g., defining a rating of 5 or more, rather than 6 or more, on a Likert scale rating as a "yes") would have produced an h-index that correlated more strongly with the well-being measures, as was found in the training data. However, a threshold of 6 was selected to best replicate the findings reported in McConnell et al. (2005).

For the most part, these findings add credence to the claim that greater self-complexity is negatively associated with well-being (Rafaeli-Mor & Steinberg, 2002). Although most empirical data suggest that greater self-complexity is negatively associated with well-being, support has been mixed due to competing evidence that greater self-complexity can also be positively associated with well-being. The data in this study supports the predominant claim, although the correlations were fairly small in magnitude. Self-aspect positivity, on the other hand, correlated much more strongly with the well-being measures, suggesting that some combination of self-complexity and self-

aspect positivity may best predict well-being, a prediction in line with the spillover hypothesis (Linville, 1985). Future work could explore this idea further by examining the degree to which self-aspect positivity moderates the relationships between self-complexity and well-being or the degree to which self-complexity predicts these well-being measures above and beyond self-aspect positivity. However, the focus of the current study was not to clarify the relationship between self-complexity and well-being, but rather to use well-being as outcome variable for checking the effectiveness of the replication.

An obvious limitation of the study is that the study only serves as a conceptual, rather than a direct, replication. In doing so, the study applies an old calculation (the h-statistic) to a new method, in which participants are asked to rate self-aspects traits on matter of degree rather than sorting traits into distinct self-aspect categories. Saying that a trait describes a self-aspect as a "6" out of "7", one could argue, is not psychologically the same as selecting a trait to be characteristic of a self-aspect in a card sort task. However, self-complexity, as calculated in this study, was only meant as an approximate replication of previous work, an interim step in the development of a new measure.

It's still worth considering, though, that the experience of completing this new, adapted, online version of the self-aspect task is psychologically different from completing the original card sort task. A traditional card sort task is a constructive, bottom-up process for determining the self-aspects that best describe an individual. The process often takes about 25 minutes (Showers, 1992), and self-aspect labels are not generated until after the traits are sorted. Alternatively, in this adapted version of the task, participants are asked to create their self-aspect labels *before* making trait ratings and are

only required to spend a short amount of time (2 minutes) doing so. Participants are then, somewhat tediously, asked to rate each self-aspect on every one of the possible traits, resulting in a less-deliberated and potentially less accurate set of ratings for each self-aspect. To some extent, this task structure was merely a consequence of the available online methodologies. Future online tasks could conceivably use an online version of a card sort task before engaging in the trait ratings, but the optimal strategy for eliciting self-aspects remains to be determined. Regardless of any practical limitations, an online version of the self-aspect task has a number of possible implications for expanding the size and types of populations that can be studied. Making the measure broadly accessibly might also be an effective way to help individuals learn about themselves (e.g., via websites like MySocialBrain, https://mysocialbrain.org/).

Future work developing the task for broader use may want to consider the trait words included in the task. The current study used the same list of traits as McConnell et al. 2005, a list of traits used as common self-descriptors amongst college students (Showers, 1992). However, because these traits serve as the basis for measuring self-aspect overlap, the degree to which these trait words can accurately characterize a given self-aspect or the relationships between any given set of self-aspects, particularly outside of an academic community, should be examined. The nature of the relationships among the self-aspects - derived using the same traits words but analyzed using modern methods - is explored in Study 2.

Study 2: Extension

Method

Open science. The design, hypotheses, and analysis plan for this study were pre-registered at the Open Science Framework (https://osf.io/t3hvj/). In this study, the results of Study 1 were extended to include novel network analysis techniques. Information about the participants, design, and procedures for this study are identical to those described in Study 1. 789 participants (403 female, *Mage* = 38.29, *SD* = 12.06) generated a list of their self-aspects and made trait ratings for each.

Analysis plan. The primary analysis sought to replicate the results of Study 1 using network analysis techniques. Trait ratings between each participant's self-aspects were correlated to generate correlation matrices. The R package "qgraph" (Epskamp, Cramer, Waldrop, Schmittman, & Borsboom, 2012) used these correlation matrices to calculate centrality measures (strength, betweenness, closeness, and expected influence) for each self-aspect for each participant. Centrality values for each self-aspect were averaged for each person to create an overall centrality measure. Different measures of centrality and clustering were tested on the test sample. Measures of particular interest included mean strength (the mean weighted summation of absolute value edge lengths for each self-aspect), mean expected influence (the mean weighted summation of non-absolute value edge lengths for each self-aspect), and mean closeness (the reciprocal of the sum of the shortest paths between each self-aspect and all other self-aspects). The measures that most closely replicated the results from Study 1 (determined through iterative testing) were selected, while also considering the shape of the distribution and the variance for each option. The selected network measures were then applied to the hold-out sample, and the resulting network metrics (i.e., centrality measures), as well as the correlation between those metrics and the well-being measures are reported.

Results

I hypothesized that network metrics (e.g., centrality) would capture similar information about self-complexity as the h-statistic, while at the same time yielding more variance and thus more information to correlate with other psychological constructs (e.g., well-being measures). Different centrality measures were tested in the training sample before one was selected for use in the test sample. Figures 2 and 3 illustrate examples networks from participants who have high and low levels of centrality, respectively.

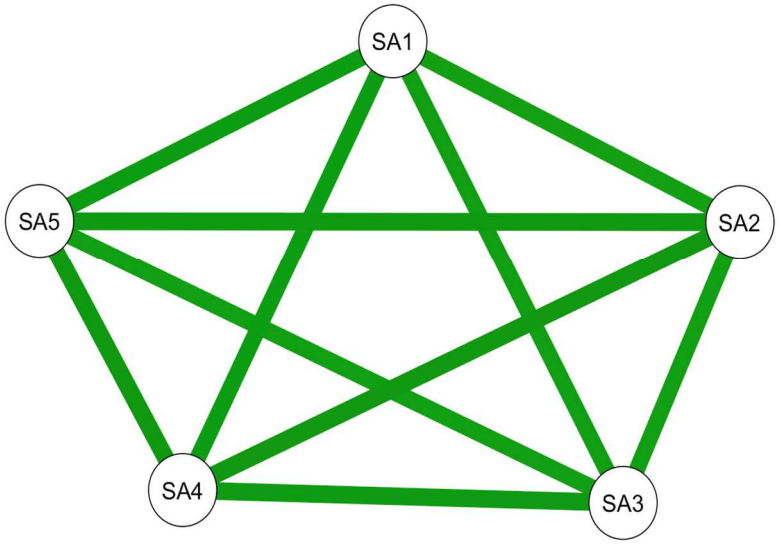

Figure 2. Example self-aspect network from a participant who has five self-aspects high in centrality. Green lines represent positive correlations, and the thick lines represent strong correlations between self-aspects.

Consistent with the theoretical assumptions of the multiple self-aspects framework, edge thresholding was set at zero to include all self-aspect correlations. Mean closeness (M = 0.20, SD = 0.15), strength (M = 2.50, SD = 1.40), and expected influence (M = 2.27, SD = 1.53) across the self-aspects for each participant from the training sample were selected for correlation with the h-index and the well-being measures from Study 1. Betweenness was eliminated as a possible metric in that it did not yield enough

variance (perhaps due to the small size of the networks). Other network-level metrics, such as clustering, were not ultimately tested given the relatively small size of the networks (although these metrics may be interesting for individuals with a higher number of self-aspects).

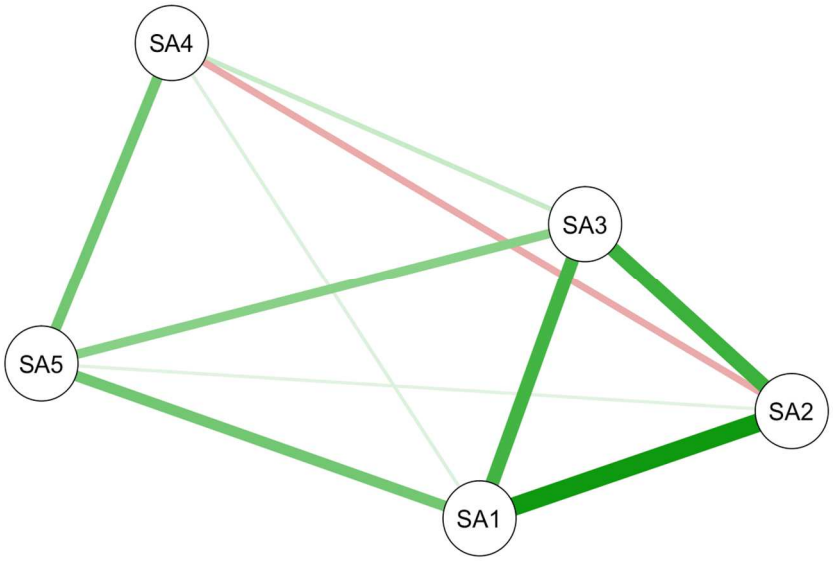

Figure 3. Example self-aspect network from a participant who has five self-aspects lower in centrality. Green lines represent positive correlations, and red lines represent negative correlations. Thickness of the line indicates strength of the

Table 2 reveals that closeness was negatively correlated with strength and expected influence, which shared a strong positive correlation with one another. Closeness, strength, and expected influence correlated with the h-statistic to varying degrees. Even so, closeness, strength, and expected influence were positively correlated with self-positivity and self-esteem, and negatively correlated with depression, perceived stress, and physical symptoms. Expected influence demonstrated the overall strongest (highest magnitude) correlations with these well-being measures.

Notably, an exploratory analysis revealed that the number of self-aspects did not significantly correlate with any of the well-being measures (Table 2), challenging

previous assumptions that self-complexity should weight both the number of self-aspects as well as the degree of overlap in its calculation (at least for the purposes of correlating with well-being). To determine the effect of controlling for the number of self-aspects (n), a linear transformation (dividing by n-1) was applied to the expected influence measure for each individual, resulting in the equivalent of a mean distance calculation (M = .58, SD = .31). Indeed, this measure correlated even more strongly with the outcome measures, with higher mean distance correlating negatively and strongly with desirable well-being outcomes such as self-esteem and correlating positively and strongly with undesirable well-being outcomes such as depression and stress (Table 3). This result suggests that network measures predict outcomes of interest above and beyond simply the number of self-aspects in a network. Table 3 also reveals that mean distance was particularly strongly correlated with self-aspect positivity. However, a regression model indicated that mean distance significantly predicted measures like self-esteem (R^2 *change* = 5.74% ; $F(489) = 40.12$, $p < .001$) and depression (R^2 *change* = 7.61% ; $F(489) = 86.64$, $p < .001$) above and beyond self-aspect positivity. Therefore, mean distance was selected as the centrality measure of interest for examination in the test sample.

Mean distance calculated for each participant in the test sample (M = .56, SD = .32) was correlated with the h-index and the well-being measures from Study 1 (Table 4). Like in the training sample, mean distance was negatively correlated with the h-index, strongly positively correlated with desirable well-being measures like self-esteem and strongly negatively correlated with undesirable well-being measures like depression, physical symptoms, and stress. Notably, mean distance was particularly strongly positively correlated with self-complexity, but a regression model revealed, once again,

that mean distance significantly predicted well-being measures like self-esteem (R^2 *change* = 7.93%, $F(296) = 38.38$, p <.001) and depression (R^2 *change* = 8.34%, $F(297) = 91.88$, p <.001) above and beyond self-aspect positivity.

Table 2
Intercorrelations among number of self-aspects and well-being measures from Study 2 training data

	No. Self-Aspects
Self-aspect positivity	0.03
Self-esteem	0.07
Depression	0.01
Physical Symptoms	0.05
Perceived Stress	-0.04
Self-aspect negativity	0.03
Self-aspect centrality	0.04
Self-concept clarity	-0.01

Note. *$p < .05$, **$p<.01$.

Table 3
Intercorrelations among self-complexity (H), centrality measures, and well-being measures from Study 2 training data

	Zero-order correlations				
	Closeness	Strength	Expected Influence	Mean Distance	*H*
Centrality: Closeness	-				
Centrality: Strength	-.23**				
Centrality: Influence	-.12**	.94**	-		
Centrality: Distance	.44**	.54**	.72**	-	
Self-complexity (*H*)	-.54**	.20**	0.07	-.30**	-
Self-aspect positivity	.19**	.40**	.51**	.63**	-.12**
Self-esteem	0.09	.34**	.43**	.50**	-.12**
Depression	-.12**	-.26**	-.36**	-0.46**	0.23**
Physical Symptoms	-.11*	-.21**	-.25**	-.35**	.22**
Perceived Stress	-.10*	-.31**	-.39**	-.46**	.16**
Self-aspect negativity	-.24**	-.40**	-.52**	-.67**	.26**
Self-aspect centrality	.25**	.33**	.38**	.72**	-0.07
Self-concept clarity	.15**	.28**	.36**	.47**	-.18**

Note. *$p < .05$, **$p<.01$.

Table 4
Intercorrelations Among Self-Complexity (H), Centrality Measures, and Well-Being Measures from Study 2 Test Data

	Zero-order correlations		
	Mean Distance	*H*	Pos.
Self-complexity (*H*)	-.20**	-	-
Self-aspect positivity	.70**	-.12**	-
Self-esteem	.59**	-.12**	.56**
Depression	-.55**	0.23**	-.49**
Physical Symptoms	-.37**	.22**	-.26**
Perceived Stress	-.54**	.16**	-.49**
Self-aspect negativity	-.70**	.26**	-.86**
Self-aspect centrality	.50**	-0.07	.65**
Self-concept clarity	.53**	-.18**	.49**

Note. *$p < .05$, **$p<.01$.

Discussion

I hypothesized that the findings of Study 1, in which greater self-complexity was negatively correlated with desirable measures of well-being and positively correlated with undesirable measure of well-being, could also be replicated using network-based approaches. Indeed, mean centrality measures across self-aspects, including mean closeness, mean strength, and mean expected influence within the training data all exhibited this pattern (although notably, because of the way that these measures are calculated, the direction of the patterns were reversed such that higher centrality, or more overlap, was associated with more well-being). Mean strength and expected influence correlated more strongly with well-being measures than closeness as a result of the way they weighted information about the number of self-aspects. Mean strength and mean expected influence also correlated more strongly with well-being measures than the h-

statistic, suggesting these measures capture meaningful information that h-statistic does not, at least with regards to well-being.

Interestingly, the number of self-aspects did not correlate with these well-being measures, indicating that a novel measure (at least one optimized for correlating with well-being) need not incorporate the number of self-aspects. As a result, average path length (mean distance), which averages out information about the number of self-aspects, was selected as an optimal measure for this analysis, and, as predicted, correlated most strongly with well-being measures. Moreover, despite the fact that average path length (mean distance) was strongly correlated with self-aspect positivity, it contributed unique information for predicting the well-being measures above and beyond self-aspect positivity.

Overall, then, the selected measure of interest in this study, the mean distance, seems to capture important information about the relationships between people's different self-aspects (and not about the number of self-aspects). The information that is captured seems to be highly related to well-being, such that individuals with higher scores (i.e., more integrated self-aspects), report experiencing more well-being. The measure is also strongly related to self-aspect positivity, but captures additional information beyond positivity.

These findings highlight the ability of network metrics to capture important information and extend the previous literature in a number of meaningful ways. Whereas previous calculations of self-complexity (using the h-statistic) have weighted both the number of self-aspects and their overlap as important factors for characterizing self-complexity, the literature has not yet had the tools available to determine which of these

component factors is related to well-being. Of course, the number of self-aspects might be meaningful for relationships with other variables, beyond well-being. The beauty of a network approach, though, is that different network metrics can be pulled from the network depending on the question of interest. For example, although the current study was interested in pulling a single network-level metric of integration to correlate with measures related to well-being, another study may simply be interested in pulling node-level information about an individual's "work self" and "home self" to examine relationships between these self-aspects and outcomes related to productivity and time management.

Although it's possible that the number of listed self-aspects may matter in other contexts, it's important to think about why they did not matter here. At the start of the task, participants are asked to provide anywhere from 2 to 8 self-aspects, and it's possible that participants choose the number of self-aspects that they would like to list somewhat arbitrarily. Whether two self-aspects are deemed to be one and the same is a rather subjective and arbitrary decision. For example, someone could choose to either list their "wife" and "mother" self-aspects separately or to lump them together as a "family" self-aspect. While more meaningful information may be obtained from listing the self-aspects separately, it may not make sense to call this person more complex simply for listing more aspects. In other words, the number of self-aspects may not provide as much meaningful information as the relationships between them. The role of the number of self-aspects will need to be further examined, but if, in fact, the number of self-aspects does not matter in meaningful ways, this finding would have a number of important implications for psychologists. First, this finding would lend more support to the idea that

psychologists studying the structure of self-aspects should move away from using measures like the h-index that are driven in part by number of self-aspects and towards a measure that is agnostic to them. Moreover, finding that the number of self-aspects does not matter would allow the number of self-aspects to be set across participants in future work without losing much meaningful information. Controlling for the number of self-aspects (and perhaps even content) would allow for much tighter comparisons across participants without needing to control for the number of self-aspects listed.

Given that the perceived similarity between the self-aspects (i.e., the degree to which participants rated self-aspects similarly across the traits), rather than the number of self-aspects listed, seems to be predictive of well-being, it's worth thinking about the type of information that this measure captures. For example, integration might, to a large degree, capture positive information about the self-aspects. Based on the spillover hypothesis (Linville, 1985), positive self-aspects are more likely to integrate, whereas negative self-aspects are more likely to disperse. However, despite the strong correlation between integration (mean distance) and positivity, integration predicted measures of well-being above and beyond positivity. In fact, the same results hold when negativity is added to the model. Thus, integration must have some other non-affectively based effect (e.g., inducing a feeling of wholeness or a feeling of control) that future work may want to investigate.

Of course, although the present study was limited to primarily examining the association between self-complexity and well-being, well-being is not the only outcome of interest. Self-complexity may be related to a whole other suite of processes. For example, the results from this study indicate that integration is positively related to self-

concept clarity, and it's possible that integration is associated with other identity-based individual difference measures. The variance for the centrality measures was relatively high, increasing the likelihood that these measures will correlate with other potential measures of interest.

The study was also limited in that the selected centrality measure, mean distance, only yields one number, essentially a mean correlation value across all self-aspects, for the entire network, washing out a lot of potentially important information about the relationships between the self-aspects. Specifically, the same integration value (mean distance) might differ in meaningful different ways across two individuals. While one person may have consistently moderate interactions between all of their self-aspects, another may have a few very strong connections in combination with a number of weak, or even negative correlations, averaging out to an overall moderate level of integration. These two very different structures might, however, yield two very different relationships with well-being or might be very differentially influenced by interventions. The strength of a network-based approach is that it is well-suited to address these types of questions, even though they were not directly addressed here. Study 3 addresses some of these question by investigating the consistency of integration over time, as well as its susceptibility to manipulation / intervention.

Study 3: Manipulation

Method

Open science. The design, hypotheses, and analysis plan for this study were pre-registered at the Open Science Framework (https://osf.io/t3hvj/).

Participants. Three weeks after participation, the same 789 participants from Study 1 were invited to participate in a 30-minute online follow-up study using TurkPrime's data acquisition platform (Litman, Robinson, & Abberbock, 2017). Participants were excluded if they did not complete the survey (either voluntarily or for failing an initial attention check), or failed either of two other attention checks included in the survey. The final sample for analysis included 520 participants (275 female, *Mage* = 38.89, *SD* = 12.10).

Design. Participants were randomly assigned to one of two writing conditions (integration or control) before completing the trait ratings for their self-aspects, yielding a mixed within (time), between (condition) subjects design.

Procedures. After consenting, participants were randomly assigned to one of two conditions that both involved a writing task. In the experimental condition, participants wrote about the ways in which their different self-aspects (generated in the first session and provided to them in a listed format here) come together and fit with one another to form their overall identity. To clarify the instructions, participants in both conditions were presented with example writing responses for another individual (Harry Potter). In the control condition, participants wrote about the details of their typical day. Writing about a typical day has been used as a control condition in other writing manipulations (e.g., Peters, Flink, Boersma, & Linton, 2010) because it controls for writing and self-focus but does not induce the specific kind of self-focused thought as the experimental condition. Participants were required to brainstorm about their prompt for at least one minute and to write for at least 5 minutes. Moreover, participants were required to write

at least 100 words before proceeding to the next section (examples responses from participants in each condition are included in Appendix B)

Afterwards, participants were provided with a list of the self-aspects they had generated in the first session and completed trait-level and overall-level ratings for each self-aspect that they listed in Study 1. Participants were asked "to what extent does the following word describe your _____ self-aspect?" (1 = not at all to 7 = very much) for the same 40 trait words used in Study 1. Presentation of the questions was blocked so that participants answered all 40 questions for one self-aspect before moving on to the next self-aspect, with order of self-aspect blocks randomized. At the end of each block, participants were asked three additional questions about overall centrality, positivity, and negativity for each self-aspect.

At the end of the survey, participants completed a post-questionnaire in which they indicated whether the self-aspects they saw in this study were the same self-aspects they listed in Study 1 (yes, no, and not sure), and provided some free responses indicating whether they would change any of the self-aspects they had originally listed, remove any of the self-aspects they had originally listed, or add any new self-aspects. Upon completion, participants were presented with a debriefing form describing the study and a code to submit to MTurk in exchange for payment. Given that some participants listed more self-aspects than average and took longer to complete the survey, participants who took longer than the estimated time were provided additional bonus payment for their time (based on the same hourly rate).

Analysis plan. The primary analyses investigated the effects of time and manipulation. A mixed factorial ANOVA was run in order to determine main effects of

time and of manipulation, as well as their interaction, with follow up contrasts to clarify any effects. Follow up contrasts examining changes from time 1 to time 2 in the control condition and examining differences between the control and integration condition at time 2 were of particular (*a priori*) interest. The same analyses were run using the h-statistic, centrality measures, mean overall self-reported centrality, mean overall self-reported positivity, and mean overall self-reported negativity as dependent measures.

Results

Based on the efficacy of previous self-based writing manipulations, I hypothesized that the writing manipulation would change the way that individuals perceive themselves, and that network metrics would be particularly sensitive to these changes. Moreover, based on findings in the personality literature that self-reported personality items should remain consistent over time, particularly over the short time period of a few weeks (e.g., Caspi, Roberts, & Shiner, 2004), I hypothesized that no significant differences in self-complexity (as measured by the h-statistic or by network metrics) would be observed between time 1 and time 2 for those participants in the control condition.

Self-complexity (h). A mixed factorial ANOVA investigating self-complexity, as measured by the h-index, revealed a significant main effect of time ($F(1, 518) = 13.61$, $p < .001$), such that self-complexity decreased from time 1 ($M = 2.33$, $SD = .79$) to time 2 ($M = 2.25$, $SD = .80$) across both conditions, but revealed no significant main effect of condition ($F(1, 518) = .08$, $p = .78$) and no significant interaction ($F(1, 518) = 0.53$, $p = .47$).

Follow-up contrasts examining the main effect of time revealed that the h-statistic (i.e., self-complexity) decreased in both the control (time 1 $M = 2.32$, $SD = .74$; time 2 $M = 2.23$, $SD = .79$; $t(259) = -3.37$, $p < .001$) and experimental groups (time 1 $M = 2.33$, $SD = .83$; time 2 $M = 2.26$, $SD = .82$; $t(259) = -1.96$, $p = .05$). Contrasts comparing differences in the integration condition ($M = 2.33$, $SD = .83$) and the control condition ($M = 2.32$, $SD = .74$) at time 1 ($t(518) = -.03$, $p = .98$) and the integration condition ($M = 2.26$, $SD = .82$) and the control condition ($M = 2.23$, $SD = .79$) at time 2 ($t(518) = -.48$, $p = .63$) were not significant.

Centrality (mean distance). A mixed factorial ANOVA investigating the primary centrality measure of interest (mean distance) revealed a significant main effect of time, such that centrality increased from time 1 ($M = .59$, $SD = .31$) to time 2 ($M = .61$, $SD = .31$; $F(1, 518) = 5.14$, $p = .02$) overall. There was no significant main effect of condition ($F(1, 518) = .12$, $p = .73$) and no significant interaction ($F(1, 518) = 0.53$, $p = .47$).

Follow-up contrasts examining the main effect of time revealed that centrality increased in the experimental condition (time 1 $M = .59$, $SD = .31$; time 2 $M = .62$, $SD = .31$; $t(259) = 2.81$, $p < .01$) but not the control condition (time 1 $M = .59$, $SD = .30$; time 2 $M = .60$, $SD = .32$; $t(259) = .60$, $p = .54$). Figure 4 illustrates network graphs for a participant who increased in mean distance from time 1 to time 2. Contrasts comparing differences in the integration condition ($M = .59$, $SD = .31$) and the control condition ($M = .59$, $SD = .30$) at time 1 ($t(518) = .06$, $p = .95$) and the integration condition ($M = .62$, $SD = .31$) and the control condition ($M = .60$, $SD = .32$) at time 2 ($t(518) = -.73$, $p = .47$) were not significant.

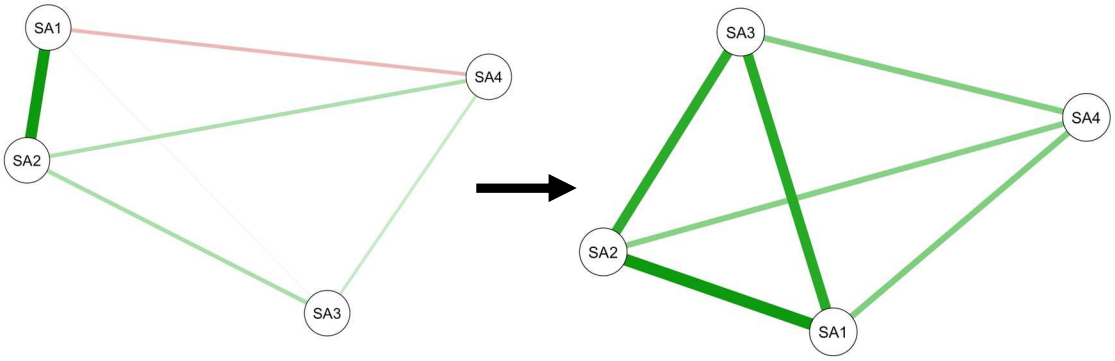

Figure 4. Example self-aspect networks from a participant who listed four self-aspects and who increased in mean distance from time 1 (left, distance = .17) to time 2 (right, distance = .54). Green lines represent positive correlations, and red lines represent negative correlations. Thickness of the line indicates strength of the relationship.

Centrality (expected influence). A mixed factorial ANOVA investigating a secondary centrality measure (mean expected influence) revealed a significant main effect of time such that self-complexity increased in general (time 1 $M = 2.28$, $SD = 1.52$; time 2 $M = 2.36$, $SD = 1.56$; $F(1, 518) = 5.14$, $p < .05$), but there was no significant main effect of condition ($F(1, 518) = .12$, $p = .73$) and no significant interaction ($F(1, 518) = 0.53$, $p = .47$).

Follow-up contrasts unpacking the main effect of time revealed that expected influence increased over time in the experimental group (time 1 $M = 2.28$, $SD = 1.46$; time 2 $M = 2.39$, $SD = 1.52$; ($t(259) = 3.14$, $p < .01$) but not in the control group (time 1 $M = 2.29$, $SD = 1.57$; time 2 $M = 2.33$, $SD = 1.61$; $t(259) = .93$, $p = .36$). Contrasts comparing differences in the integration condition ($M = 2.28$, $SD = 1.46$) and the control condition ($M = 2.29$, $SD = 1.57$) at time 1 ($t(518) = .12$, $p = .90$) and the integration condition ($M = 2.39$, $SD = 1.52$) and the control condition ($M = 2.33$, $SD = 1.61$) at time 2 ($t(518) = -.38$, $p = .70$) were not significant.

Self-reported centrality. A mixed factorial ANOVA investigating self-reported self-aspect centrality revealed no significant main effect of time ($F(1, 518) = 2.44$, $p = .12$), no significant main effect of condition ($F(1, 518) = 1.92$, $p = .26$), and no significant interaction ($F(1, 518) = 1.85$, $p = .18$).

Follow-up contrasts (for consistency) revealed that mean self-reported self-aspect centrality did not change over time in the experimental group (time 1 $M = 5.55$, $SD = .91$; $M = 5.54$, $SD = .90$) ($t(259) = -.15$, $p = .88$), although showed a trending but non-significant (Bonferonni-corrected) decrease in the control group (time 1 $M = 5.68$, $SD = .95$; time 2 $M = 5.59$, $SD = 1.02$) ($t(259) = -2.03$, $p = .04$). Contrasts comparing differences in the integration condition ($M = 5.55$, $SD = .91$) and the control condition ($M = 5.68$, $SD = .95$) at time 1 ($t(518) = 1.60$, $p = .11$) and the integration condition ($M = 5.54$, $SD = .90$) and the control condition ($M = 5.59$, $SD = 1.02$) at time 2 ($t(518) = .48$, $p = .63$) were not significant.

Self-aspect positivity. A mixed factorial ANOVA investigating mean self-reported self-aspect positivity revealed no significant main effect of time ($F(1, 518) = .27$, $p = .61$), no significant main effect of condition ($F(1, 518) = .06$, $p = .81$), but did reveal a significant interaction ($F(1, 518) = 6.10$, $p < .05$) (Figure 5).

Follow-up contrasts examining the interaction revealed a cross-over effect such that self-reported self-aspect positivity increased (non-significantly) in the integration group (time 1 $M = 5.80$, $SD = .95$; $M = 5.86$, $SD = 1.00$) ($t(259) = 1.45$, $p = .15$), and showed a trending but non-significant (Bonferonni-corrected) decrease in the control group (time 1 $M = 5.89$, $SD = 1.03$, time 2 $M = 5.81$, $SD = 1.03$) ($t(259) = -2.02$, $p = .04$). Contrasts comparing differences in the integration condition ($M = 5.80$, $SD = .95$) and the

control condition (M = 5.89, SD = 1.03) at time 1 (t(518) = .97, p = .33) and the integration condition (M = 5.86, SD = 1.00) and the control condition (M = 5.81, SD = 1.03) at time 2 (t(518) = -.49, p = .62) were not significant.

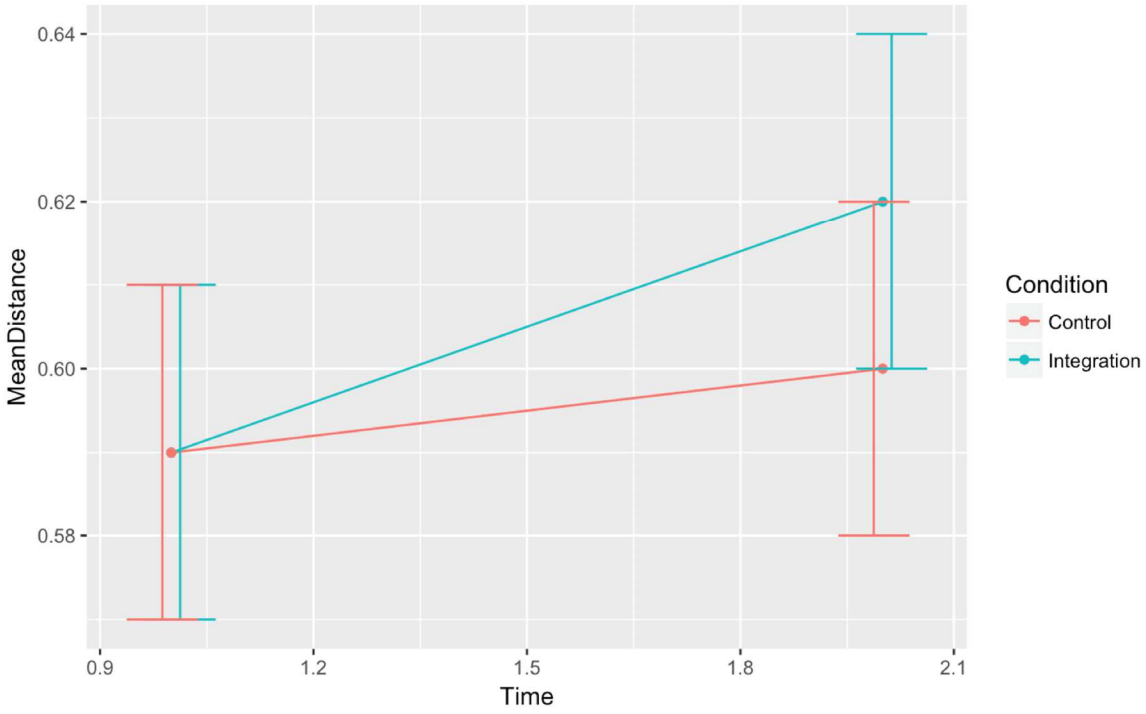

Figure 5. Interaction effect for self-aspect positivity across time for the two conditions (control and self-integration).

Self-aspect negativity. A mixed factorial ANOVA investigating mean self-reported self-aspect negativity revealed no significant main effect of time F(1, 518) = .16, p = .69, no significant main effect of condition (F(1, 518) = .64, p = .42), and no significant interaction (F(1, 518) = 0.18, p = .69).

Follow-up contrasts examining the interaction revealed a cross-over effect such that self-reported self-aspect negativity did not change in the integration group (time 1 M = 1.98, SD = .98; M = 1.96, SD = 1.02) (t(259) = -.63, p = .53) or in the control group (time 1 M = 2.04, SD = 1.09, time 2 M = 2.04, SD = 1.09) (t(259) = .02, p = .98). Contrasts comparing differences in the integration condition (M = 1.98, SD = .98) and the

control condition (M = 2.04, SD= 1.09) at time 1 (t(518) = .64, p = .52) and the integration condition (M = 1.96, SD = 1.02) and the control condition (M = 2.04, SD = 1.09) at time 2 (t(518) = .88, p = .38) were not significant.

Discussion

I hypothesized that a self-integration manipulation would lead to increases in measures of self-complexity and centrality compared to a control condition. I tested this prediction across two studies using different operationalizations, with self-complexity indexed by the h-index in Study 1 and centrality by mean distance and expected influence in Study 2. For exploratory purposes, the effects of the manipulations on self-aspect positivity, negativity, and self-reported centrality were also investigated. A change over time was found for both self-complexity and centrality such that, surprisingly, self-complexity significantly decreased in the control condition and, less surprisingly, centrality significantly increased for the integration condition. Moreover, self-aspect positivity decreased for the control condition but increased for the integration condition, suggesting a possible mechanism for the effects on self-complexity and centrality reported above. Although these findings were, for the most part, in line with the hypothesis, they were not found in the predicted ways (e.g., although the integration condition changed across time, it was not significantly different from the control condition).

The somewhat similar findings for self-complexity (using the h-statistic) and complexity (using network-based approached) lend credence to the idea that both are, to some extent, measuring similar information, although it's unclear which is more reliable for detecting the effects of intervention. Notably, the change over time for self-

complexity was driven by the control condition, whereas the change over time for centrality was driven by the integration condition, suggesting that the two measures are capturing different types of information, likely as a result of the way they deal with self-aspects or counting responses (as binary or as a matter of degree). Network approaches have the overall advantage of being more flexible, calculating metrics differently based on the needs of the question, but the best measure for capturing changes over time has yet to be determined, as mean distance and expected influence were optimized for correlation with well-being measures, not detecting change over time.

The interaction effect for positivity suggests that affect may play an important role in self-complexity changes. In this case, integration promoted increases in positivity for the self-aspect ratings, whereas the control condition promoted decreases in positivity for the self-aspect ratings. Based on these results, it's possible that thinking about the ways that selves come together can be affirming or mood-boosting, in some way. Indeed, the integration manipulation was a new manipulation designed for this particular study, and any underlying mechanisms for engaging in the task are unknown. Similarly, although the control condition was intended to be neutral, it's possible that thinking about more typical elements of one's life could be somewhat mood deflating.

Although the results hint that the integration manipulation may lead to (small) changes in the relationships between self-aspects, the fact that both conditions led to changes on measures of either self-complexity or self-integration suggests that self-aspects, regardless of any manipulation, change over time. In other words, the effects of life, itself, may change self-aspect ratings more so than the effect of a 5-minute writing exercise. Still interesting is the idea that self-aspect ratings changed in similar directions

across participants, in favor of decreased complexity and higher integration. In this sense, it's possible that the first session actually served as a longer-lasting manipulation, priming participants to start thinking about their self-aspects over the next 3 weeks in ways they had not thought of them before, or perhaps simply inducing some sort of habituation at time 2. These possible reasons for the change observed are entirely speculative. Regardless of the underlying reasons, though, it is worth underscoring that the relationships between people's self-aspects did significantly change over time, whereas no change would have been expected by chance.

That said, the integration manipulation was limited in a number of ways. First, although the manipulation was grounded in previous literature, the integration manipulation was a new manipulation designed for the purposes of this study. Other more established manipulations could have been more effective at manipulating self-integration, and those aimed at decreasing rather than increasing integration should be explored in future work. Even so, this particular manipulation was chosen to avoid any potential third variables associated with other, established manipulations. The goal of the present study was to increase self-integration, and so a manipulation asking participants to list the ways that their self-aspects are integrated was developed. An advantage of this approach is that it avoids certain questions about the mechanisms underlying the intervention. However, a potential disadvantage of the approach is that there may have been certain demands characteristics in telling people to write about their integrated self-aspects and then having them rate the degree of integration amongst their self-aspects. Given that integration was not directly measured (only trait ratings), it is not expected

that any demand characteristics would have been obvious to the participants, but the limitation is still worth considering.

Even so, the finding that identity changes across both conditions holds implications, both for the broader literature on identity change (e.g., Caspi, Roberts & Shiner, 2005) as well as for future intervention-based work. Because identity is usually deemed to be stable, evidence of identity-based changes lends support to the idea that identity can be used as a target for intervention. Whereas transformative life events are "special" in that they yield quick, large-scale changes in identity (McAdams, 2008; McAdams & Guo, 2015), intervention may be able to achieve similar effects through smaller-scale, sustained changes. Future work will need to not only explore other interventions that might lead to change in self-complexity, but also examine the degree to which these types of self-complexity interventions are long-lasting over time. The network methods examined in this study provide one potential method for measuring those changes, as well as for comparing the mechanisms underlying existing, self-based manipulations (e.g., self-distancing, self-affirmation, self-compassion).

As part of this network-based approach, future work will also be able to consider the more nuanced relationships between self-aspects. Although a measure of overall integration may capture important information about a person, as a whole, it may not be the most important level at which to measure the relationships between self-aspects in the case of interventions. Rather than targeting the self as a whole, interventions may want to target particular self-aspects, for example, by integrating a self that does not correlate strongly with the rest of the self-aspects. Future work should consider the content of the particular self-aspects listed (e.g., using a factor analysis) and the effects that

manipulations have on particular self-aspect relationships, a relationship that would become easier to examine in the case of standardizing self-aspects across participants.

General Discussion

Knowns

Through a series of experiments, this paper replicated the relation between previously established self-complexity measures and well-being, extended these measures to the domain of network analysis, and showed that these measures of self-complexity change over time. Although conclusions regarding the relationship between self-complexity and well-being in the literature have been mixed, the strength of the correlations between self-complexity and well-being measure found in this large sample of data supports the idea that individuals with lower self-complexity tend to experience better well-being. Importantly, network measures of self-complexity were, by a large margin, able to explain even more of the variance in well-being, further supporting the idea that individuals with lower self-complexity tend to experience better well-being and demonstrating one way in which network approaches can reveal important information about the structure of the self. Lastly, although no one manipulation yielded more change in self-complexity over time, overall change in self-complexity across time was found using a number of measures. The knowledge that change is possible will be informative for future interventions that aim to directly manipulate self-complexity.

Unknowns and Future Directions

Together, these known findings evoke broader questions about the nature of the self and highlight the continued need to develop tools for measuring self and identity. First, the findings evoke broader questions about the nature of the self and whether or not people can be said to have one self or many. Simply because individuals are able to

report multiple important self-aspects does not mean that they perceive themselves this way on a daily basis. Rather, the task required that participants list more than one self, and many participants may have simply been meeting these task demands. However, it is likely that some individuals do experience having multiple selves, and yet this does not mean that they do not also have a single, unified sense of self. Indeed, these self-aspects were, on average, fairly strongly positively correlated with one another, indicating some sort of unity amongst them. Rather than asking *whether* participants are able to hold both self-perspectives, though, research in the field asks *how* individuals are able to maintain both multiple self-aspects and a unified sense of self (Roberts & Donahue, 1994). Whether these individual self-aspects can be said to comprise that same overall sense of self or whether the sense of self is something entirely different remains a philosophical question.

Second, the findings evoke broader questions about identity change (explored more in Chapter 3). Researchers examining life narrative acknowledge that the types of transformative life events that change the way individuals talk about themselves occur, by definition, infrequently over the course of a lifetime (McAdams, 2008; McAdams & Guo, 2015). Moreover, personality researchers note that an individual's identity-defining traits can change, but that these types of changes occur over the course of decades. To what extent, then, can a brief intervention or manipulation be said to lead to any form of identity change, and do significant changes in calculated self-complexity or centrality qualify as identity change? These questions, too, are philosophical in nature, but the answers have the potential to inform directions for future research. For example, rather than attempting to manipulate identity in small ways, researchers might try calculating

changes in self-complexity and centrality reported by individuals experiencing transformative life events.

Finally, the findings provoke a broader call towards the development of methods for studying the self. Although network-based behavioral approaches provide a novel method for investigating questions about the structure of the self, this method will need to be validated against other measures. Of particular interest are measures that are able to better clarify the broad mechanisms underlying self-complexity and the multiple self-aspects framework. For example, it is currently unclear whether each self-aspect takes up a separate cognitive "workspace" that becomes differentially activated, or whether each self-aspect "comes online" in a shared self-processing workspace, a la the working self-concept (Markus & Kunda, 1986). Neuroimaging provides a valuable tool for not only answering these more process-based questions at the neural level, but also for answering these questions at the psychological level (Berridge, 1995), and is anticipated to be informative in both ways here, as well. The first step to addressing this more process-based question is to simply understand how multiple self-aspects are represented at a neural level. Although traditional studies investigating neuroscience of the self have applied univariate metrics to identify brain regions that are involved in tracking self-referential information more broadly (Kelley et al., 2002; Moore, Merchant, Kahn, & Pfeifer, 2014), more nuanced brain decoding techniques have shown that refinement within these regions is possible (e.g., Chavez, Heatherton, & Wagner, 2016; Yankouskaya, Humphreys, Stolte, Stokes, Moradi, & Sui, 2017). Determining the extent to which similar decoding techniques detect multiple self-aspects within pre-identified

regions of interest (e.g., vmPFC) and align with the behavioral network metrics describe above is a promising next direction.

CHAPTER III

PERSONAL IDENTITY AND THE MORAL SELF

What is Personal Identity?

Imagine a person, "J," who lives in the Pacific Northwest, has an affinity for outdoor activities, tends to be introverted, and has unwavering interest in the self. Fast forward 30 years to a scene in which J still lives in the Pacific Northwest, hikes when she can, still lives a relatively quiet life, but no longer studies the self in any way, shape, or form. Is "J" the same person she was 30 years ago?

The study of personal identity has historically centered around these and other related questions concerning the continuity of self over time and the qualities of self that determine that continuity. Philosophical perspectives on this line of work have varied, ranging from Hume, who somewhat mindfully argued that because the self is derived from fleeting "impressions," there is no way for the self to be persistent through time (1738/1978), to Parfit who invokes images of amoeba-like selves and tele transportation to Mars to argue that connectedness amongst selves is only possible as a matter of degree (1971). Albeit insightful, these topics entertained by Hume and Parfit, do not say much about who "J" is. Rather, these perspectives primarily fall under the branch of personal identity known as re-identification, which aims to characterize the type of subject that can track persistence over time (Schechtman, 1996) and is primarily of sole interest to philosophers.

The concept of re-identification (described above) stands in subtle contrast to the concept of categorization, which, rather than characterizing the subjects that track persist

over time, aims to identify the *content* that is essential to the persistence of these subjects over time (Schechtman, 1996). For example, John Locke famously espoused that psychological continuity is at the heart of personal identity, claiming that elements of consciousness, and particularly memory, are required to deem a person the same from time 1 to time 2 (Locke, 1690/2009). According to Locke, "J" would still be the same person in 30 years so long as she *remembered* that she once had a passion for studying the self, and, under this argument, it could be said that memories are "essential" to who J is. However, if "J" were instead perceived to be a different person in 30 years because of her lost interest in the self, then her interest in the self, rather than her memories, would be deemed a more essential part of her identity. Given the types of conclusions drawn from these types of thought experiments about the relative importance of certain content to identity, this branch of personal identity concerned with categorization has a number of important implications for psychologists.

Even so, psychologists have not traditionally considered the types of claims and arguments provided by philosophers in this domain. Rather, psychologists have traditionally used the term "personal identity" to reference the ways in which individuals characterize themselves as distinct from other individuals (Olson, 2016). This type of personal identity has also been called "synchronic personal identity," that domain of personal identity focused on identifying the traits and roles that characterize someone at a particular time. In the case of "J," a psychologist would reference *all* of "J's" qualities, including her identity as a Pacific Northwesterner, as an outdoorswoman, as an introvert, and as a researcher interested in the self to describe J's personal identity.

As a result, the majority of empirical attention has been dedicated towards understanding this synchronic element of self, to the exclusion of the "diachronic" self, or the self that persists over time (Northoff, 2017). The diachronic self has been almost entirely overlooked by researchers despite its clear importance to how people think about themselves and others. Of course, studies investigating elements of mental time travel (Schachter, Addis, Hassabis, Martin, Spreng, & Szpunar, 2012) and child development (Pfeifer et al., 2013) come close to addressing important elements of self-continuity, but none of the paradigms used in these studies have directly considered those core elements of personal identity that persist over time.

The study of self as narrative gets close and has even been argued to comprise the core of who we are (Dennett, 1992), but philosophers have pointed out that there is a difference between the narrative self and personal identity. Whereas narrative is a collection of the identities that one possesses, personal identity references the core subject that possesses all of those identities (Peacocke, 2014). Narrative "J" would be the person who possesses the consistent identities over time (the Pacific Northwesterner, the outdoorswoman, and the introvert), whereas arguments from personal identity would say that "J," at her core, is a person who researches the self because she deems herself a different person otherwise.

All of these perspectives on personal identity provide valuable information in the pursuit of effectively characterizing and understanding an individual. Whereas some perspectives, such as the traditional psychological perspective, have long been a topic of thorough empirical scrutiny, other perspectives, such as the traditional philosophical perspective, have not. That said, understanding the content that is at the core of

someone's identity is invaluable for fully understanding that person. Here, I focus on the empirical investigation of personal identity as traditionally defined in the philosophical tradition; in particular, I focus on the approach of categorization that attempts to identify the content that is essential to a person.

What Elements of Identity Are Essential?

Empirical philosophers have started identifying the content that is essential to a person by gathering lay perceptions on the matter. Original experiments utilized body-switched scenarios in order to test folk intuitions about identity. Blok, Newman, and Rips (2005) asked participants to imagine that the brain of one person ("J") had been placed into the body of another person. In one scenario, memories are not preserved such that J, despite maintaining her cognitive faculties, does not remember who she is when in the other body. In another, memories are preserved, such that she does remember who she is. People tend to agree that J is the same person only when memories are preserved, an effect that has been replicated across a variety of scenarios (Nichols & Bruno, 2010), supporting the conclusions of many philosophers dating back to Locke.

However, the notion that memories are most core to who we are has recently been challenged. Strohminger & Nichols (2014) asked participants to imagine a variety of science-fiction-like examples in which one part of someone changes, but everything else remains the same. For example, participants were told to imagine that someone had taken a pill that would selectively change only one part of that person's mind but nothing else and were then asked to indicate the degree to which that someone would be a different person given a particular change. Across five separate studies, more than changes to perceptions (e.g., ability to feel pain), desires (e.g., enjoyment of a favorite food),

memories (e.g., cherished memories of time spent with parents) and personality (e.g., industrious), changes to morality (e.g., honesty) yielded most perceived change, to a relatively strong degree.

This "moral self effect" has since been replicated across a variety of studies (e.g., Heiphetz, Strohminger & Young, 2017; Everett, Skorburg, Livingston, Chituc, & Crockett, under review), including evidence indicating that the effect holds even in the real world (Strohminger & Nichols, 2015) and across different cultures (Garfield, Nichols, Rai, & Strohminger, 2015). Moreover, the finding that morality is core to who we are is consistent with work in separate but related areas of investigation in empirical philosophy. A line of work investigating the nature of the "true" self finds that that the self, at its core, is not only moral, but also good (Newman, Bloom, & Knobe, 2014), including evidence that moral gains lead to less perceived identity change than moral losses (Tobia, 2016). Moreover, elements of morality have been implicated in a line of work investigating the "deep" self, which argues that identifying the true self has important implications for assigning moral responsibility (Sripada, 2016).

What Is Special About Morality?

This growing line of evidence highlights the importance of morals to human identity. However, it is still not clear why moral traits are perceived as essential. Although definitions of morality are vast, morals are broadly defined as an agreed upon set of norms within a society and are distinguished from other social norms primarily in the sense that they are more important or more valued (Hare, 1952). At their core, then, morals are inherently social (albeit in a special way), and any changes to morality are likely to impact more than one individual within a community (Heiphetz, Strohminger, &

Young, 2016). Specifically, Person A is likely to care about a change to the morality of Person B because Person A will no longer receive the benefits of being socially tied to Person B. Indeed, morality is a key dimension on which we perceive other individuals (Goodwin et al., 2014). However, Person A is also likely to care about her own morality given that it impacts her social reputation.

Implicit in this social dynamic is the idea moral traits are, in either case (change to other or change to self), quite important, or valuable, to the self. Indeed, Heiphetz, Strohminger, & Young (2016) found that the importance of traits to the self mediated judgments of identity change, with moral traits rated as more important, suggesting some role for value-based processing in moral perception. This line of reasoning is further supported by increasing evidence that value-based reasoning plays an important role in moral decision-making (Bench-Capon & Modgil, 2017). Although the deontological tradition within philosophy has long assumed that moral decision-making relies on logical reasoning, evidence increasingly suggests that reward and value-based processing drive moral decision-making (Shenhav & Greene, 2010; Shenhav & Greene, 2014). Specifically, when participants are asked to imagine a scenario in which they can either do nothing and risk the death of a large group of people or do something at the expense of a single individual, value-based sub-regions of the brain are largely involved in making these types of calculations, suggesting that moral reasoning, to a large degree, is driven by value-based calculation.

Although mounting evidence suggests that morals tend to be both socially-oriented and value-based, there are many ways in which a trait can fit both of these criterion, and it is worth considering how moral traits differ from other types of traits. For

example, an individual might value the interactions they have with a stranger because that stranger was warm and friendly towards them. Moreover, an individual might value the interaction they have with a stranger because that stranger demonstrated competence in an area and was helpful to them. Warmth and competence provide two key dimensions upon which people perceive one another (Fiske, Cuddy, & Glick, 2007). Although some moral traits are certainly lower in warmth and higher in competence (e.g., temperance or prudence), it is typically the warmer moral traits (e.g., honesty, loyalty) that most readily impact social preferences (Leach & Barreto, 2007) and social information gathering (Brambilla, Rusconi, Sacchi, & Cherubini, 2011). Therefore, it is hypothesized that warmer moral traits are more influential for the moral self effect, and that some sort of value-based processing drives it, but mechanisms underlying the moral self effect have yet to be explored.

Practical Implications

Clarifying the mechanisms underlying judgments of personal identity is important not only for furthering broader understanding of self, but it may also provide relevant insights to more translational work. The identity-value model of self-regulation states that identity serves as a salient value-input for facilitating successful self-regulation, and that stable, value-laden sources of identity are strongest (Berkman, Livingston, & Kahn, 2017) (Chapter 1). If morality is, in fact, core to identity and is driven by value-based processing, it may be a candidate target for interventions seeking to promote behavior change. Of course, given that morality is so essential to identity, it may also be tougher to manipulate than other aspects of identity, but the evidence suggests to the contrary.

Indeed, appealing to moral identity can motivate moral behavior (e.g., Hardy & Carlo, 2005; Hardy & Carlo, 2011). Moreover, appealing to moral reasoning has been shown to motivate compliance on certain behaviors such as paying taxes (e.g., Blumenthal, M., Christian, C., Slemrod, J., & Smith, M. G, 2001; Ariel, 2012) and environmental conservation (Bolderdijk, Steg, Geller, Lehman, Postume, 2012; Hopper & Nielson, 1991). Given that many self-regulatory failures are often moralized (Rozin & Singh, 1999; Frank & Nagel, 2017), appealing to moral identities and values may also be an effective strategy for motivating successful self-regulation, as well.

Counterfactual thought experiments, such as those traditionally used by philosophers, might also play a key role in motivating self-regulation. Many effective self-regulation techniques already draw upon hypothetical and imaginative cognitive techniques encouraging individuals to think about themselves in new and alternative ways (Kross et al., 2014; White et al., 2016). Encouraging participants to imagine the degree to which they would become a new person if they were to achieve a goal (e.g., "become a whole new you!") could provide an avenue for examining ways in which identity and value facilitate self-regulation. Exploring the mechanisms underlying the traditional moral self effect, although not directly related to translational applications, may be able to help motivate work in this direction.

In fact, real world applications of the moral self effect are already being explored. For example, Strohminger & Nichols (2015) demonstrate that families of loved ones struggling with frontotemporal dementia (which impairs moral faculties), do actually rate these individuals as being different people (compared to pre-disease) more than families of individuals struggling with other health conditions, suggesting that moral changes can,

in fact, impact identity perceptions in the real world. These perceptions about the effects of disease may extend to perceptions about the effects of addiction, as well. A recent study found that people perceived an addicted person to be a different when compared to their non-addicted self, and that these perceptions are driven by moralistic judgments of the addiction (Earp, Skorburg, Everett, & Savulescu, under review). However, much more work remains to be done in this area.

First and Third-Person Asymmetries

If the findings of the moral self effect are, in fact, going to be extended to have real-world implications, the role of target considered for the judgment of identity change must be considered. Specifically, if morality is going to be used as an identity-based motivator, the interventions would be applied in the first-person. However, personal identity, as examined in the traditional sense, is a third-person study of self, designed to draw conclusions about personhood, more broadly.

Traditional findings in social psychology examining person perception show pervasive effect of target, such that perceptions of self are often biased in certain ways in comparison to perceptions of another (and vice versa). The actor-observer bias demonstrates that although people are more likely to ascribe positive events to internal causes and negative events to external causes for themsselves, the opposite is true when judging others (Ross, 1977; Ross, Amabile, & Steinmetz, 1977). To a large extent, this asymmetry is a consequence of positive biases associated with the self: people generally tend to rate themselves above average than others on most things (Taylor & Brown, 1988, Sedikides, Gregg, & Toguchi, 2003). However, these biases are also simply influenced by the sources of knowledge we have about the self: in general, we more accurately

evaluate internal information about ourselves given our privileged access to it, whereas we are more accurate in evaluating external information about others, again, given the privileged access (Vazire & Carlson, 2010).

Given that these first and third-person asymmetries are so pervasive in the field of psychology, similar findings would be expected within the field of personal identity. Moreover, if morality is fundamental for social reasons (as hypothesized above), these asymmetries would be expected to persist for judgments of identity change: people are expected to weight identity change differentially for self and for others given that they do not interact with themselves in the same way that they interact with others. However, to date, first and third-person asymmetries have not been observed for studies investigating identity change. One of the first studies to consider the role of target showed that a body-switching paradigm yielded similar results regardless of target (Nichols & Bruno, 2010). More recently, no difference was observed in a series of studies directly comparing the moral self effect for self and a hypothetical other ("Chris"), albeit showing stronger effects for other compared to self on certain moral traits (Heiphetz, Strohminger, & Young, 2016). Another series of studies examining the effect of target across many different categories showed that the moral self effect holds across self, known friend, and unknown stranger, but not for a known enemy (although the self and friend condition were not compared directly), again showing a small to negligible effect of target for self and other.

It has been hypothesized that the lack of asymmetry between self and other reflects the implicit positive nature of the moral self across most targets (Strohminger, Newman, & Knobe, 2016). Indeed, the true self literature has demonstrated that although

our own true selves are deemed to be inherently good, so are the true selves of others (Bench et al., 2015). This lack of affective bias for moral traits may be one reason that the effect is not observed for cases of personal identity. However, there are other plausible reasons. Although there are well-known asymmetries for perceiving self and other, there are also asymmetries and biases for perceiving the self in the past, present, and the future (Ersner-Hershfield, Wimmer, & Knutson, 2008; Quoidbach, Gilbert, & Wilson, 2013), such that thinking about the self in a hypothetical thought experiment may not be the same as thinking about the self in the here and now. Moreover, many of the targets in the above studies were not specified others. Given that judgments of moral identity change may be driven by values, the values might change if a different, more concrete target were used, and actual perceptions of the target might influence the results. Regardless of whether the behavioral effects change in different circumstances, the mechanisms driving the effects, or lack thereof, remain to be explored.

Underlying Brain Mechanisms

In sum, there are two types of mechanisms underlying the moral self effect that remain to be clarified. First, it is unclear why morality is special. Although it has hypothesized that morality is important for social reasons, the cognitive architecture underlying that "specialness" remains to be determined, with valuation as a plausible mechanism. Second, it is unclear why there is not an effect of target for judgments of personal identity. Even if a behavioral effect is found, it is unclear whether thinking about personal identity relies on similar mechanisms as engaging in traditional thought about self and other. Neuroscience can be a particularly helpful tool for dissociating the

mechanisms underlying a phenomenon (Berridge, 1996). The brain regions hypothesized to underlie these distinct elements of the moral self effect are reviewed below.

Self and Other-Perception

Neuroimaging studies investigating self-referential processing (in the traditional psychological sense) have primarily investigated the effects of trait-based perception, asking individuals to indicate whether a series of trait word describe themselves as well as another individual. This trait-based processing tends to recruit the cortical midline structures of the brain (precuneus and medial prefrontal cortex), with self-referential activity localized in the more ventral region of the medial prefrontal cortex and other-referential processing localized in the more dorsal region of the medial prefrontal cortex (Denney et al., 2012; Wagner et al., 2012).

Notably, these cortical midline regions of the brain are also involved in a whole suite of other types of cognitive processing, and there is debate as to what exactly these brain regions are tracking, especially given that information about the self (as compared to another person) varies on a variety of important dimensions. For example, as reviewed earlier in this chapter, information about the self tends to be inherently more positive than information about other individuals, and unsurprisingly, this difference is also reflected at the neural level, with self-referential neural activity sharing highly overlapping patterns with both positively valenced information (Chavez, Heatherton, & Wagner, 2017) and overlapping activity in regions of the brain that are closely tied to reward and value-based processing (Berkman, Livingston, & Kahn, 2017; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). In fact, many have even claimed that the self is "nothing but reward" (Northoff & Hayes, 2011) or personal relevance (Kim & Johnson, 2015).

Information about the self, by its very nature, is also more familiar than information about other individuals, and brain regions tracking self-relevant information may be tied not only to value, but also to familiarity (Lin, Horner, & Burgess, 2016)

Regardless of what is being tracked in these regions, the relative constellation of activity can be particularly informative, particularly along the ventral to dorsal gradient. Indeed, studies have shown that the closeness of a target to the self can be tracked such that targets closer to the self activate more ventral regions of the medial prefrontal cortex whereas targets less close to the self activate more dorsal regions of the medial prefrontal cortex (Mitchell, Banaji, & Macrae, 2006), with very close targets demonstrating overlapping activity (Zhu et al., 2007). Moreover, the activity within these regions varies, to some extent, with elements of hypothetical thought. Counterfactual thinking for self and for other has demonstrated a similar ventral to dorsal dichotomy in the brain (DeBrigard, Spreng, Mitchell, and Schacter, 2015). Thinking about the self in the future, too, tends to recruit regions of the cortical midline structures, as part of the default mode network (Buckner & Carroll, 2007; Buckner, Andrews-Hanna, & Schacter, 2008). Whereas some studies have found that thinking about the self in the future activates more dorsal (as compared to ventral) regions of the medial prefrontal cortex (Packer & Cunningham, 2009), other have found that thinking about the self in the future compared to the present simply recruits the ventral medial prefrontal cortex to a lesser degree (Ersner-Hershfield, Wimmer, & Knutson, 2008, Tamir & Mitchell, 2011; D'Argembeau et al., 2010). However, no study, to present knowledge, has investigated the type of hypothetical thought required for making identity change judgments.

It is unclear, then, whether neural activity associated with identity change judgments will show similar patterns of activity as those associated with traditional self and other-perception. On the one hand, these patterns are well-established and reliably elicited. However, judgments of personal identity do not show the same behavioral self-other asymmetries as traditional self and other-perception in the social psychological literature, suggesting that thinking about personal identity for self and other might recruit more overlapping neural processes than is traditionally observed. At a general level, given that thinking about personal identity is very person-centered, activation in cortical midline structures of the brain is likely still expected. More specifically, if self-other asymmetries are absent from the personal identity literature because thinking about a hypothetical self is akin to thinking about another person, overlapping activity in dorsal regions of the medial prefrontal cortex might be expected. Alternatively, if self-other asymmetries are absent from the personal identity literature because thinking about another individual changing on a moral dimension is valuable to the self, overlapping activity in ventral regions of the medial prefrontal cortex might be expected. In either case, overlapping activity would likely suggest that thinking about personal identity is less target-centered than traditional self and other perception, lending support to the philosophical assumption that studying third-person personhood and studying first-person narrative are, indeed, two separate branches of self-study.

Morality and Value

The brain mechanisms underlying processing of different trait categories are less well understood. Aside from differences in positive and negative trait information (Glisky, & Marquine, 2009; Fossati et al., 2004), few studies have reported any neural

differences in processing the domain of a trait (Pfeifer et al., 2013; Pfeifer et al., 2009). One of the reasons why differences across traits are not reported is simply because effects of traits using traditional univariate analysis techniques are simply not found. However, more nuanced multivariate techniques are revealing differences in the neural representations of the types of information (e.g., rationality, social impact, and valence) that organizes our mental states for people (Tamir, Thornton, Contreras, Mitchell, 2016).

Given that moral traits are deemed to be more essential to personhood than other types of information, it is hypothesized that some underlying mechanism would track that information. Although no studies to date have investigated the contribution of moral traits to self and other-perception, per se, many studies have investigated the neural mechanisms involved in moral reasoning. Perhaps most famously, in contrast to deontological arguments assuming that moral judgments require logical reasoning, recent neuroimaging evidence suggests that moral reasoning tends to recruit activity in value and reward-based regions of the brain (Shenhav & Greene, 2010; Shenhav & Greene, 2014). Moreover, a recent meta-analysis found that tasks requiring moral judgment (vs. non-moral judgment) heavily recruited regions of the default mode network, leading the author to conclude that moral reasoning is largely intertwined with self-based processing (Han, 2017). Other recent work found that morals recruit more dorsal regions of medial prefrontal cortex, leading the authors to conclude that morals may be more akin to basic social cognition (Theriault, Waytz, Heiphetz, & Young, 2017).

Based on somewhat sparse evidence, it is not entirely clear which brain regions might be involved in supporting the moral self effect. Given that morals are hypothesized to be important to self and that moral reasoning seems to heavily rely upon value-based

processing, it perhaps seems most plausible that moral traits will also recruit value-based regions of the brain, either selectively or simply more strongly, than other non-moral traits. Alternatively, given the hypothesized social role of morality, it is also possible that moral traits, relative to non-moral traits, will recruit more activity in dorsal regions of the medial prefrontal cortex. In maintaining these hypotheses, it is important to keep in mind that differences between moral and non-moral traits may not be detectable at a univariate level and may require subtler, multivariate methods.

Present Study

The following suite of studies aimed to extend the work on personal identity to better clarify the mechanism listed above, namely: 1) Why is morality special? and 2) Is thinking about personal identity for different targets similar to engaging in traditional person perception (and if so, why)? Given the hypothesized role of values in the moral self effect, I aimed to remove as many potential confounds in trait words as possible. Study 1 piloted and matched words on a variety of important dimensions, including valence in preparation for Study 2, which sought to replicate the basic moral self effect (Strohminger & Nichols, 2014) while controlling trait words for valence. Although these studies were exploratory in nature, it was hypothesized that trait words could be matched for valence and, given the large effect sizes reported in Strohminer & Nichols (2014) that the effect would replicate. Study 3 aimed to extend the moral self effect by comparing first and third-person perceptions of identity change. The moral self effect was predicted to replicate, and, despite the fact that self-other asymmetries are typically observed within social psychology, no main effect of target was hypothesized (Heiphetz, Strohminger, & Young, 2016). Finally, Study 4 used fMRI to examine the mechanisms underlying

thinking about personal identity for self and for other, as well as the mechanisms underlying the basic moral self effect. Again, the study was somewhat exploratory, but based on previous literature, it was hypothesized that a main effect of target would reveal differential activity in ventral and dorsal regions of the brain for self and other-based judgments of identity change, respectively, and that a main effect of trait would reveal stronger activity in the ventromedial prefrontal cortex for moral versus non-moral traits. No significant interaction effect was hypothesized.

Study 1: Trait Development

Method

Participants. 351 participants (18 years of age or older; native English speaker; U.S. resident) were recruited from the Amazon Mechanical Turk population using TurkPrime's data acquisition platform (Litman, Robinson, & Abberbock, 2017) to participate in a 15-minute online study rating different kinds of traits. Sample size was determined through a power analysis using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) showing that a sample of 54 observations per trait would be required to detect a moderate effect size (d = .5) with 95% power using a two-tailed, one-sample t-test. Participants were excluded if they did not complete the survey, failed either of the two included attention checks, or completed the survey in less than 2 minutes. The final sample for analysis included 334 participants (129 female, $Mage$ = 34.49, SD = 10.87), yielding the suggested number of observations per trait (M = 59.06, SD = 1.69).

Design. Participants were randomly assigned to view and rate a total of 100 trait words (out of 281 possible trait words) for either its relevance to morality or its degree of

positivity. Trait words were selected to best represent the full range of possible moral (Strohminger & Nichols, 2015) and positively-valenced trait words (Anderson, 1968).

Procedure. After completing the pre-screen and the consent form, participants completed a demographic form indicating their age, gender, ethnicity, and highest level of education. Participants assigned to rate traits for morality were instructed that they will see a series of traits and, for each, were asked "Is this trait related to morality?" (1 = Not at all related to morality, 7 = Extremely related to morality). Participants assigned to rate traits for valence were instructed that they will see a series of traits and, for each, were asked "How positive is this trait?" (1 = Neutral, 7 = Extremely Positive). Participants in both conditions were shown 100 traits words, presented randomly with 10 traits per page. At the end of the survey, participants were presented with a code to submit to MTurk in exchange for payment.

Analysis plan. The primary analysis was to determine whether moral and non-moral traits could be matched for valence. To do this, one-sample t-tests were run for each trait word in order to determine which traits were significantly different from the midpoint of the morality scale (4 = Somewhat related to morality). Of the traits that were significantly higher than the midpoint, the 40 trait words rated as being most relevant to morality were selected. Of the traits that were significantly lower than the midpoint, the 40 trait words rated as most positive were selected. An independent sample t-test (equal variance not assumed) was run to determine whether these two lists of words differed in valence.

Results

I hypothesized that moral trait words could be matched with non-moral trait words on the dimension of valence. An independent sample t-test comparing the valence for the 40 trait words rated as being most related to morality (M = 5.72, SD = 0.50) and the 40 trait words that, of the words rated significantly lower than the midpoint on the morality scale were rated most highly for valence (M = 5.62, SD = 0.25), yielded no statistically significant difference ($t(58.07) = 1.21$, $p = .23$, demonstrating that moral and non-moral traits can, in fact, be matched for valence. Of note, as a result of the trait selection method, the variance for valence was larger for the moral trait words, and equal variance between the conditions was not assumed.

Study 2: Replication

Method

Participants. 137 participants (18 years of age or older; native English speaker; U.S. resident) were recruited from the Amazon Mechanical Turk population using TurkPrime's data acquisition platform (Litman, Robinson, & Abberbock, 2017) to participate in a 15-minute online study about personality changes. Sample size was determined through a power analysis using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) showing that a sample size of 90 would be required to detect a small to moderate effect size (d = .3) (as reported for changes to moral versus personality traits in Study 3 from Strohminger & Nichols, 2014) with 80% power using a two-tailed, paired-sample t-test. Participants were excluded if they did not complete the survey, failed either of the two included attention checks, or failed one of the two task comprehension questions. The final sample for analysis included 109 participants (52 female, $Mage$ = 35.33, SD = 11.51).

Design. Participants were asked to make identity change judgments for both moral and non-moral traits (matched for valence, as determined in Study 1) in a within-subjects design.

Procedure. After completing the pre-screen and the consent form, participants completed a demographic form indicating their age, gender, ethnicity, and highest level of education. In the next section, drawing on the instructions from Study 2 from Strohminger & Nichols (2014), participants were asked to imagine that scientists have developed pills that, once swallowed, would permanently alter only one part of someone's mind, without affecting anything else. This particular vignette was selected in order to avoid the temporal components present in many of the other vignettes. In the questions that followed, participants were instructed to imagine that someone took one of these pills such that everything about them exactly the same, except they are no longer X, with direction of change (decrease) specified drawing on previous work indicating that losses are treated differently than gains (Tobia, 2016). The scenario was presented separately for each of the 40 moral and 40 non-moral traits (selected from Study 1a), yielding 80 total questions, presented in randomized order with 10 traits per page. For each question, participants were asked "Do you agree they are still the same person as before?" (1 = Strongly disagree, 5 = Strongly agree). At the end of the survey, participants were asked to complete some task comprehension questions, some free response questions about the types of traits that led to more perceived identity change and about the nature of the target (i.e., the "someone") that they thought about, questions about political orientation, the Moral Identity Scale (Aquino & Reed, 2002), and the

Implicit Person Theory Scale (Chiu, Hong, & Dweck,1997). Finally, participants were presented with a code to submit to MTurk in exchange for payment.

Measures. The primary dependent measure looked at identity change: to what extent would someone be a different person after experiencing the change? The scale for this dependent measure was adapted from the 1-100 point sliding scale originally used by Strohminger & Nichols (2014) into a 5-point scale in preparation for the fMRI experiment (Study 4). Of note, this scale was anchored such that lower scores would indicate more change, whereas Study 3 and 4 were anchored such that higher scores would indicate more change.

Analysis plan. The primary analysis was to determine whether the effect between moral and non-moral traits originally reported by Strohminger & Nichols (2014) would replicate when matched for valence. A paired sample t-test between identity change ratings for moral and non-moral traits was used to test this question.

Moreover, the results from this study were used to select 20 moral and 20 non-moral traits for use in Study 3 and Study 4. Traits were selected to optimize for moral traits that most yielded perceived identity change relative to non-moral traits (i.e., optimizing for the "moral self effect") while still controlling for valence, length, syllables, and frequency in the English language, as determined by comparing words from each group on an independent sample t-test.

Results

I hypothesized that the findings of Strohminger & Nichols (2014) would replicate such that perceptions of identity change for moral traits would be stronger than perceptions of identity change for non-moral traits, even when controlling for valence. A

paired sample t-test comparing the identity change ratings (with lower scores indicating more change) for both moral (M = 3.42, SD = 0.96) and non-moral (M = 3.82, SD = 0.96) traits yielded a statistically significant replication of the original results ($t(108) = -6.14$, $p < .0001$, supporting the hypothesis that moral traits are more essential to a person's identity .

Moreover, it was hypothesized that moral trait words and non-moral trait words could be refined to 20 per category to match on dimensions pertaining to valence, length, syllables, and frequency in the English language (Brysbaert & New, 2009). A series of independent sample t-tests revealed that the selected moral (M =5.81, SD = 0.39) and non-moral (M = 5.74, SD = 0.27) words did not significantly differ on valence ($t(33.91) = 0.65$, $p = .52$); that the selected moral (M = 8.65, SD =2.56) and non-moral (M = 8.95, SD = 2.19) words did not did not significantly differ on word length ($t(38) = -0.40$, $p = .69$); and that the selected moral (M =2.34, SD =0.85) and non-moral (M = 2.40, SD = 0.86) words did not did not significantly differ on frequency in the English language ($t(38) = -0.22$, $p = .83$), demonstrating, again, that moral and non-moral trait words can be matched on a variety of. See Appendix C for a full list of moral and non-moral words selected for use in Study 3 and Study 4.

Discussion

It was hypothesized that the moral self effect would replicate even when controlling. Indeed, participants rated that changes to moral trait would lead to more perceived identity change than changes to valence-matched non-moral traits, with a similar effect size as the original moral self effect.

Previous work on the moral self effect has not controlled trait words on linguistic dimensions in any meaningful way, leaving open the possibility that valence may have been driving the original results. However, the results of this study, which used empirically driven lists of trait words, suggest valence is not a confounding variable in the moral self effect. Equally interesting is the finding that moral trait words can be matched for valence.

This set of results suggests that valence is likely not a defining feature of morality, leaving open the question, what is? The list of words used in this study, as well as in Study 3 (see Appendix C), seem to differ systematically on dimensions of warmth and competence. Indeed, a follow-up to the present study confirmed that participants rated moral trait words from this study more strongly on warmth and non-moral trait words from this study more strongly on competence (Livingston et al., in prep). These results indicate that there may not be something special about morality, per se, but something special about warmth. Whereas interacting with someone who lacks competence is a hindrance, interacting with someone who lacks warmth is more directly hurtful. Future studies may want to further probe differences between thinking about identity change judgments for traits differing on dimensions of warmth and competence.

The present study was limited in that, like previous research, it did not consider the role of target. Participants in this study were only asked to think about a general, unspecified person (i.e., "someone"), which, as discussed in the introduction, leaves the underlying mechanisms unknown. Participants instructed to think about a vague "someone" could be thinking about a specific individual as an example, could be thinking about the "average other," which introduces positivity biases, or could be thinking about

themselves and their own trait values. Study 3 probes addresses this issue by specifying the particular target (self vs friend).

Study 3: Extension

Method

Open science. The design, hypotheses, and analysis plan for this study were pre-registered (currently embargoed) at the Open Science Framework (https://osf.io/8yw2p/).

Participants. 137 participants (18 years of age or older; native English speaker; U.S. resident) were recruited from the Amazon Mechanical Turk population using TurkPrime's data acquisition platform (Litman, Robinson, & Abberbock, 2017) to participate in a 25-minute online study about identity, personality, and values. Sample size was determined through a power analysis using MorePower 6.0 (Campbell & Thompson, 2012) showing that a sample size of 74 would be required to detect a small, within-subjects interaction (d = .2) with 80% power. However, previous work (Heiphetz et al., 2016) used a sample size of N = 103 with a similar design, so this study sought to collect usable data from at least 103 participants to match the previous work. Participants were excluded if they did not complete the survey, completed the survey more than once, failed more than two of six included attention checks, failed more than one of the two condition checks, or failed more than one of the two post-experiment checks. The final sample for analysis included 123 participants (67 female, $Mage$ = 35.69, SD = 10.34).

Design. Participants were asked to make identity change judgments for both themselves and another person, across both moral and non-moral traits (selected from Study 2) in a fully within-subjects design.

Procedure. After completing the pre-screen and the consent form, participants were asked to identify and type the name of a friend of the same gender and approximate age who they have known for at least one year and with whom they interact regularly (Moore et al., 2014). Identifying a friend who was familiar enough to think about but not too overlapping with the self was important in preparation for Study 4. In the next section, drawing on the instructions from Study 2, participants were asked to imagine that scientists have developed pills that, once swallowed, would permanently alter only one part of someone's mind, without affecting anything else. In the questions that followed, they were instructed to imagine that either they (themselves) or their friend took one of these pills such that everything about them is exactly the same except they are no longer X, with conditions presented in randomized order. Within each condition, the scenario was presented for each of the 20 moral and 20 non-moral traits (selected from Study 2), yielding 40 total questions, presented in randomized order with 4 traits per page. For each question, participants were asked "How much would you (yourself) change?" or "How much would ____ change?" with the friend's name piped in (1 = Same person, 5 = Different person). At the end of both conditions, participants were asked to complete some task comprehension questions, a free response questions about the types of traits that led to more perceived identity change, a free response question asking whether and why more change was perceived for themselves or for their friend, and some questions about their relationships with their friend (e.g., closeness, similarity).

After completing the questions about identity change, participants were asked to rate the extent to which each of the traits is important to them. The question was asked for each of the 20 moral and 20 non-moral traits that they considered in the previous

section, yielding 40 total questions, presented in randomized order with 10 traits per page. For each question, participants were asked "How important is this trait to you?" (1 = Not at all important, 5 = Very important). Finally, participants were asked complete a question about their political orientation, the Moral Identity Scale (Aquino & Reed, 2002), the Implicit Person Theory Scale (Chiu, Hong, & Dweck,1997), and a demographic form. Finally, participants were presented with a code to submit to MTurk in exchange for payment.

Analysis plan. The primary analysis was to determine whether the "moral self effect" (Strohminger & Nichols, 2014) would hold more strongly when thinking about oneself as compared to a friend (or vice versa), whether the effect between moral and non-moral traits would replicate when thinking about these specific targets, and whether an interaction would reveal that thinking about identity change for oneself vs. one's friend might yield stronger (or weaker) results for particular trait types. A repeated measures ANOVA was conducted to determine the main effect of target (self vs. friend), main effect of trait (moral vs. non-moral), and any potential interaction between target and trait.

Results

I hypothesized that the findings of Strohminger & Nichols (2014) would again replicate such that perceptions of identity change for moral traits would be stronger than perceptions of identity change for non-moral traits. However, based on the results of Heiphetz, Strohminger, & Young (2016), no differences between first-person (changes to self) and third-person (changes to other) judgments of identity change were predicted. Moreover, based on the results of Heiphetz, Strohminger, & Young (2016), it was

hypothesized that an interaction would be present such that participants would report stronger identity change for moral traits relative to non-moral traits when engaging in third-person vs. first-person judgments of identity change, but both of these analyses were somewhat exploratory.

A repeated measures ANOVA revealed a significant main effect of trait ($F(122) = 8.51$, $p < .01$) , such that changes to moral traits ($M = 3.65$, $SD = 1.01$) led to greater perceived identity change relative to non-moral traits ($M = 3.50$, $SD = 0.82$), as well as a significant main effect of target ($F(122) = 5.94$, $p < .05$), such that perceived identity change for self ($M = 3.65$, $SD = 0.89$) was perceived to lead to more change than perceived identity change for a friend ($M = 3.50$, $SD = 0.95$). However, the interaction effect investigating perceived identity change on moral ($M = 3.70$, $SD = 0.98$) and non-moral ($M = 3.60$, $SD = 0.79$) traits for self and on moral ($M = 3.59$, $SD = 1.04$) and non-moral ($M = 3.40$, $SD = 0.85$) traits for a friend was not significant ($F(122)= 1.37$, $p = .24$).

Discussion

I predicted that this study would, again, replicate the moral self effect. Indeed, the basic moral self effect was, again, replicated here, although the effect was somewhat smaller than that reported in Study 1. The effect of target was somewhat exploratory, and, despite the social psychological literature predicting self-other asymmetry, no effect of target was hypothesized. Interestingly, the study did reveal a main effect of target such that stronger change was reported for self as compared to other, but again, the effect was relatively small. No interaction was hypothesized, and, as predicted, no interaction effect

between target and trait was found – moral changes led to greater perceived identity change than non-moral change for both self and other.

The significant main effect of trait adds to credence to the idea that the moral self effect holds when controlling for valence. As noted, the effect was somewhat smaller in this study than reported in Study 2, and there a few reasons why this might be. First, the exclusion criteria for this study were not quite as strict for this study as they were for Study 2 – participants were allowed to fail more attention checks and were not excluded for completing the survey too fast. Applying stricter exclusion criteria might strengthen the effect, but this possibility has not been explored given the pre-registered exclusion criterion. Moreover, it is possible that the subset of traits selected for use in this study show a weaker effect than the traits used in Study 2. However, that possibility seems unlikely given that traits were selected based on their identity change ratings in Study 2. Additionally, it is possible that the conditions introduced in this study (self and other) washed out some of the effects. Participants viewed and responded to each trait word twice, which may have dampened the impact of considering some of the trait changes. Lastly, and perhaps most likely is the idea that actual perceptions of targets influenced perceptions of identity change. Although actual perceptions of targets were not collected in this study, they were collected for Study 4. Ratings of trait importance to self collected in this study could also be used to (partially) test this hypothesis. once again, that the moral self effect holds, even when controlling for valence.

The significant main effect of target suggests that traditional social-psychological phenomena influence philosophical reasoning about personal identity. Based on previous literature investigating self-other asymmetries in moral essentialism, the finding is

completely unsurprising. However, the personal identity literature has, of yet, failed to find an effect of target, a failure which philosophers have also noted as unexpected (Strohminger, 2016). One reason why the effect, albeit small, may have been observed in this study is that this is the first study to directly compare perceptions of identity change for two specific, real-world targets. Heiphetz, Strohminger, & Young investigated the effect of target but used a hypothetical other. However, comparing perceptions of identity change for self to perceptions of identity change to a hypothetical other may not have been the best approach in that the details about the hypothetical other had to be filled in by the participant, perhaps using information about the self to do so. Everett et al. (under review) investigated the effect of target and used a concrete other (a friend), and replicated the moral self effect for both, but did not test for an effect of target. One reason a concrete other may have yielded different results is that knowledge about the target may have been used to inform perceptions of identity change. For example, if a participant was asked to imagine that a friend is no longer honest, but that friend is not honest to begin with, the perceived change for that moral trait would not be very strong. If true, this would mean that individuals, on average, perceive themselves to be more moral than their friends. Given the lack of interaction, this would also mean that individuals, on average, perceive themselves to be higher on the non-moral traits, a likely finding given the literature on self-positivity biases. Unfortunately, this study did not collect data on actual self and other-perceptions, although these data were collected for Study 4.

Again, these findings suggest that social psychological biases have the potential to influence philosophical reasoning, but, of course, this would not be the first time. Experiments from social psychology demonstrating the power of the situation (e.g.,

Milgram, 1963 ; Darley & Latane, 1968) provided important challenges to moral philosophers and virtue ethicists arguing for the power of certain character traits. However, the interdisciplinary impact here is not one way. Psychologists need to be careful about the claims that can be made when adapting philosophical paradigms for their own interests. Personal identity is traditionally a third-person study of self that has been adapted here for use in the first-person. Whether personal identity adapted in this way still qualifies as the study of personal identity remains to be determined.

For now, assuming that engaging in identity change judgments about the self qualifies as a form of personal identity judgment, future work will need to tease apart the degree to which personal identity and social psychological judgments for self and other rely upon the same underlying cognitive mechanisms. However, this is tricky to do using behavioral results alone, as it is still unclear how participants are treating the information about each of the targets in Studies 2 and 3. Study 4 uses neuroimaging to clarify these underlying mechanisms.

Study 4: fMRI

Method

Open science. The design, hypotheses, and analysis plan for the behavioral pilot (Study 3) to this study were pre-registered (currently embargoed) at the Open Science Framework (https://osf.io/8yw2p/). Although the fMRI analyses reported here were not explicitly pre-registered, the behavioral pre-registration acknowledges that the study was designed with the intention of investigating the same effects of target and trait in the scanner.

Participants. 28 participants (18 female, *Mage*, = 24.57, *SD* = 4.57), with no history of psychological or neurological disorder, normal or corrected-to-normal vision, not taking medications affecting normal cognitive function, and right-handed from the University of Oregon community participated in a single-session fMRI study on perceptions of identity. Given that the task relied on subtle linguistic connotation and philosophical thinking, participants were additionally screened for native English speaking and for their undergraduate GPA (>3.0). No direct power analysis was conducted, but rather, the sample size was in line with the current recommendations for sample size in fMRI studies (Mumford & Nichols, 2008).

Design. In the identity change task (modeled off of Study 3), participants were asked to make identity change judgments for both themselves and another person across both moral and non-moral traits (selected from Study 2) in a fully within-subjects design. Moreover, a self-localizer and values-localizer were included to allow for subsequent *a priori* ROI analyses, as well as for a multi-voxel pattern analysis (MVPA) between the localizer tasks, unrelated to the current study.

Procedure. Before scanning, as in Study 3, participants were asked to identify and type into a Qualtrics survey the name of a friend of the same gender and approximate age who they have known for at least one year and with whom they interact regularly. Participants were then introduced to the pill vignette (from Study 2 & Study 3) and practiced a question thinking about themselves and a question thinking about their friend. Additionally, participants were introduced to a control condition in which they were asked to simply think about the trait itself, without any particular person in mind. Drawing from the personality literature (John & Robins, 1993) as well as control

conditions in similar studies (Moore, 2015), this control condition asked participants to evaluate the observability of the trait, considering the extent to which the trait can be easily observed in someone else (1 = Not observable, 5 = Very observable). In order to familiarize participants with the timing of all three tasks (identity change, self localizer, and values localizer), participants completed practice trials of each task prior to scanning.

All scans were acquired on a 3T Siemens Skyra Scanner at the University of Oregon's Robert and Beverly Lewis Center for Neuroimaging. Each scan session included the acquisition of epi field maps to establish any inhomogeneities in the magnetic field, a T1-weighted (MP-RAGE) anatomical image, one resting state scan consisting of echo-planar images (TR = 780ms, TE = 32 ms, matrix size = 84, 60 slices, slice thickness = 2.5 mm), and four functional runs of high-resolution, echo-planar images (BOLD-EPI) for each task (twelve runs total), collected using multiband scanning to avoid signal dropout in regions of interest (TR = 2000ms, TE = 25 ms. field of view = 208 mm, matrix size = 104, 72 slices, slice thickness = 2 mm). Stimuli were projected onto a projector, which participants viewed using a mirror placed on the head coil. Psychtoolbox was used to present stimuli and to record participant responses, which participants indicated using a five-fingered button box placed under their right hand.

After a brief scout scan localizing head position, each scan session began with a single, 15-minute resting state scan. For these scans, participants were instructed that they did not need to think about anything in particular but that they simply needed to keep their eyes open. Physiological data was collected concurrently to monitor heart rate and respiration.

The first and primary functional task investigated the brain mechanisms underlying judgments of identity change for self, other (friend), and a low-level control (observability) in an event-related design. On each trial of the task, participants were first presented with a cue indicating the type of upcoming trial ("Imagine you took a pill and are no longer…" ; "Imagine your friend took a pill and is no longer…" ; or "Think about how observable the trait is…") (3000 ms), followed by the trait adjective (moral or non-moral) paired with one of three cues ("Self," "Friend," or "Trait") (6000 ms), followed by a rating period (4000 ms) in which either the word "CHANGE?" was presented asking participants to rate the extent to which someone who lost that trait would become a different person (1 = Same person to 5 = Different person) or the word "OBSERVE" was presented asking participants to rate the extent to which the trait word is generally observable (1 = Not observable to 5 = Very observable), followed by a jittered inter-trial interval period (*M* = 4000ms sec) (Figure 6). The same forty trait adjectives (20 moral and 20 non-moral, matched for valence) from piloting (Study 3) were used in this task. Participants rated each trait adjective for each target (self and other), yielding 80 total person-specific trials. Twenty of the same trait adjectives (10 moral and 10 non-moral) were randomly selected for each participant for presentation in the control condition. Trial ordering was optimized for signal detection across the relevant domains (self vs. other & moral vs. non-moral contrasts) using an established genetic algorithm (Wager & Nichols, 2003) such that trial condition ordering was fixed across participants, although the actual traits within each category were presented in a randomized order for each participant, across four runs.

In order to directly compare neural activation associated with thinking about identity change and the traditional "self" and "other" regions of interest in future analyses, the second task consisted of an established "self" functional localizer task (Moore, 2015) in which participants judged whether 40 trait words describe them or their friend. To avoid overlap with moral and non-moral trait words, trait words that were determined to *not* be significantly different from the midpoint on the morality rating scale in Study 1 were selected for use in this task (Appendix C). To optimize self and other-referential neural activity, stimuli were presented in a blocked design, with two blocks of each condition (self and other) presented in alternating order for each run, across four runs. Each block contained five trials, and each trial consisted of an instruction period ("Describe me?" or "Describe friend?") (2000ms), combined cue (trait) and rating period (yes or no) (2000 ms), and an intertrial interval (1000ms), with blocks separated by fixation (5000 ms).

Lastly, in order to compare neural activation associated with making judgments of identity change with traditional "value-based" regions of interest (in future analyses), the final task consisted of a "willingness to pay" task commonly used in the neuroeconomics literature (Plassmann, O'Doherty, & Rangel, 2007) in which participants judged the amount that they would be willing to pay for different foods, both healthy and non-healthy. To enhance the value-signal elicited during this task, participants were instructed to refrain from eating at least two hours prior to arrival. Moreover, participants were told that at the end of the scan, their responses would be entered into an auction in which they would randomly receive one of the food items for the amount they indicated (deducted from their participant payment amount) and to only bet what they would actually be

willing to pay. Participants were presented with 16 trials (8 healthy and 8 unhealthy) for each run, repeated with new food images across four runs. Each trial consisted of a cue period presenting the food image (2500 ms), a response period asking participants to indicate how much they would be willing to pay ($0, $0.50, $1.00, or $1.50) (3000 ms), and a jittered inter-trial interval (*M* = 4500 ms).

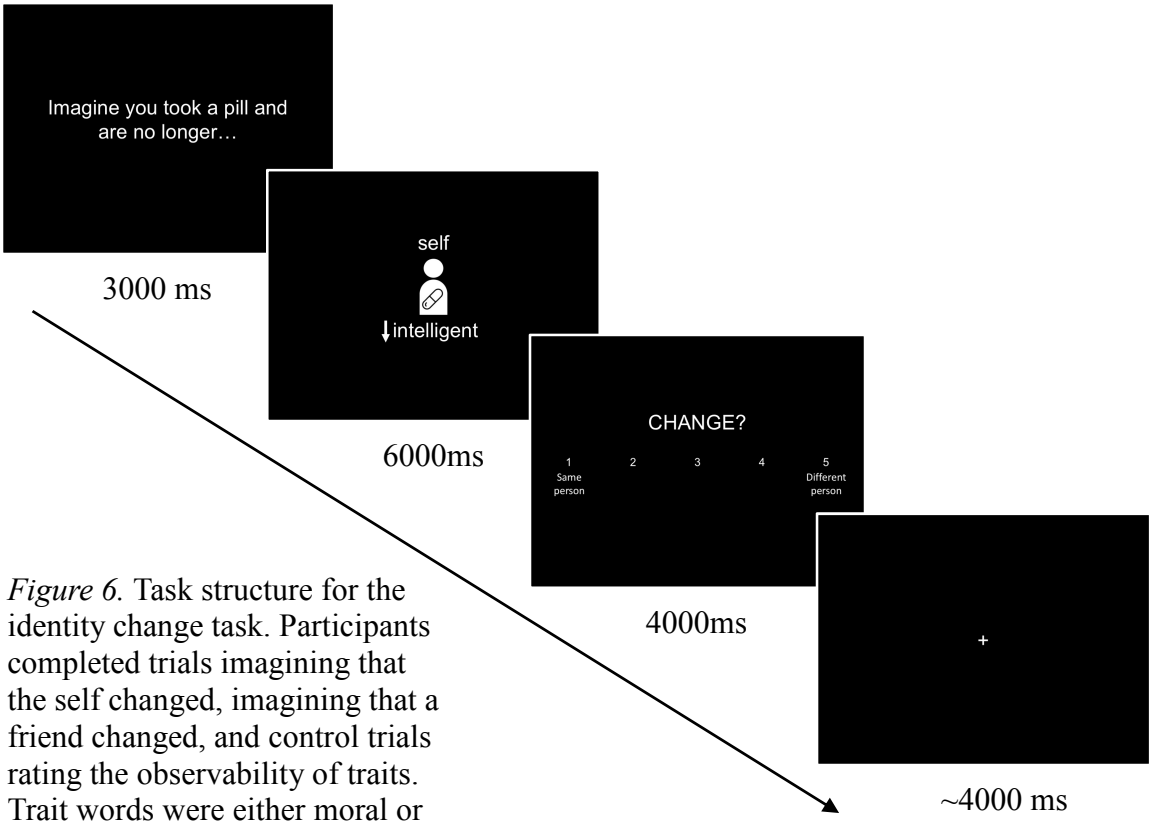

Figure 6. Task structure for the identity change task. Participants completed trials imagining that the self changed, imagining that a friend changed, and control trials rating the observability of traits. Trait words were either moral or non-moral in nature.

Upon completing the scanning portion of the study, participants completed a battery of questionnaires outside the scanner. First, participants completed some free response questions about the types of traits that led to more perceived identity change and about whether more change was perceived for themselves or for their friend. Next, participants were asked to rate the degree to which each of the trait words that they saw

in the scanner (moral and non-moral) actually describes themselves and their friend in the real world, with conditions (self and friend) presented in randomized, blocked order across participants. Within each condition, participated made ratings for each of the 20 moral and 20 non-moral traits, yielding 40 total questions, presented in randomized order with 4 traits per page. For each question, participants were asked "How much does this word describe you?" or "How much does this word describe _____?" with the friend's name piped in (1 = Not at all, 5 = Very much).

At the end of both conditions, participants were asked to complete some questions about their relationships with their friend, including how long they have known their friend, how close they feel to their friend, how similar they are to their friend, how familiar they are with their friend, and how positive they feel about their friend. Next, participants were asked to rate the extent to which each of the traits presented in the identity change scanner task is important to them. The question was asked for each of the 20 moral and 20 non-moral traits that they considered in the scanner, yielding 40 total questions, presented in randomized order with 10 traits per page. For each question, participants were asked "How important is this trait to you?" (1 = Not at all important, 5 = Very important). Then, participants were asked to consider each of the foods that they had seen in the values task and to indicate how much they like it (1 = Dislike, 4 = Love). Finally, participants were asked to indicate their political orientation and to complete the Moral Identity Scale (Aquino & Reed, 2002), the Implicit Person Theory Scale (Chiu, Hong, & Dweck,1997), the NIH-toolbox Meaning and Purpose Form (Salsman et al., 2014) (for use with the resting state data), and a demographic form. At the end of the

session, participants were given their randomly selected food item from the auction, payment, and a debriefing form.

Analysis plan. Imaging data from the project was formatted in accordance with the Brain Data Imaging Structure (BIDS) (Gorgolewski et al., 2016), the file structure used internally by OpenfMRI.org, an organization that facilitates the use of highly accessible and reproducible preprocessing streams. DICOM images were converted to NIfTI using MRIConvert software (http://lcni.uoregon.edu/~jolinda/MRIConvert/). The BIDS formatted data were preprocessed (realigned, co-registered, segmented, and normalized to the MNI template) and quality checked using fMRIprep version 1.0.12 (Esteban et al., 2017), one of the preprocessing streams available via OpenfMRI that utilizes custom code to generate advanced preprocessing streams and readable output using software from multiple neuroimaging packages. Preprocessed functional scan data outputted by the fMRIprep preprocessing stream was subsequently smoothed (with a 6mm kernel) using SPM12.

Statistical comparisons were computed using a general linear model for each participant. Activity was modeled separately with condition regressors for the instruction and cue period (with results reporting activity for the cue period, unless specified otherwise), and motion parameters estimated by fMRIprep (framewise displacement, thresholded at 0.8) entered as additional regressors. These regressors were convolved with the canonical hemodynamic response function in SPM12 and high-pass filtered with a 128 s period. This model was applied to all voxels within an explicit mask created by averaging and smoothing (6mm kernel) the structural images (gray matter and white matter) for each participant. Contrast estimate maps (e.g., self vs. other) were computed

separately for each participant, then imported to a random effects group-level analysis to estimate population-level effects. To correct for multiple comparisons, all reported contrasts were thresholded according to recommended parameters calculated by AFNI's 3dClustSim.

The primary analyses were intended to parallel those of the behavioral analyses in Study 3, investigating the neural mechanisms associated with thinking about target (self vs. other), trait (moral vs. non-moral), and their interaction. A whole-brain analysis was conducted to investigate each of these questions, comparing statistical activation maps for self and other trials (collapsed across trait), non-moral and moral traits (collapsed across target), and their interaction using one-sample t-tests. Moreover, contrasts comparing all conditions to control, a series of follow-up contrasts (reported below) were conducted for exploratory and clarification purposes. All reported coordinates are in MNI space.

Results

Main Effect of Target.

Person vs. Control. Given the person-centered nature of the identity change task, it was hypothesized that, at a basic level, thinking about either first-person or third-person identity change relative to control would activate cortical midline structures traditionally associated with person-centered thought. Statistical activation maps for all person-centered trials were compared relative to control (Figure 7). As predicted, cortical midline structures were activated for this thinking about person-level identity change relative to control, including the posterior cingulate cortex (PCC), medial prefrontal cortex (mPFC), and ventromedial prefrontal cortex (vmPFC). Notably, the medial

prefrontal activity was generally located at the more anterior end of the prefrontal cortex. Peaks for these activated regions are listed in Table 5.

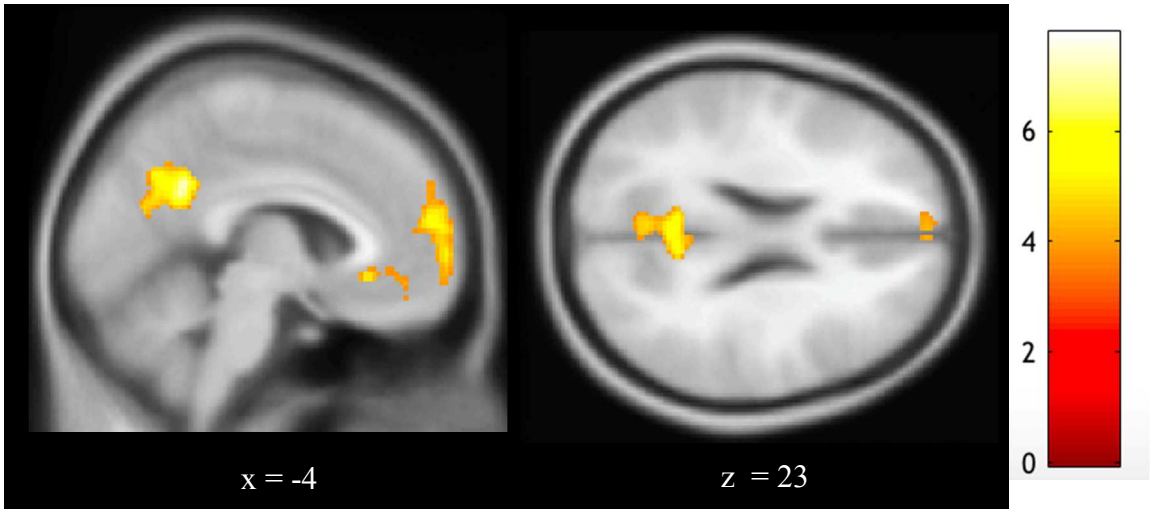

Figure 7. BOLD activity associated with making identity change judgments for all targets relative to control. Heat map refers to t values. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold

Table 5
Identification of BOLD Signal Increases for All Identity Change Trials Relative to Control

Neural Region (MNI Coordinates)	*x*	*y*	*z*	No. of voxels	Peak T
Posterior Cingulate Cortex	2	-56	38	920	7.14
Right Occipital Lobe	24	-100	12	921	6.42
Medial Prefrontal Cortex	-2	58	16	491	5.52
OFC / vmPFC	4	50	-18	202	4.77

Note. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 157. OFC = orbitofrontal cortex, vmPFC = ventromedial prefrontal cortex.

Control vs. Person. Conversely, the control task activated a series of brain regions compared to person-centered trials (Figure 8), including bilateral cerebellum, bilateral prefrontal cortex / inferior frontal gyrus (IFG) (with stronger left

lateralization), left superior temporal gyrus, and pre-supplementary motor area (pre-SMA). Peaks for these activated regions are listed in Table 6.

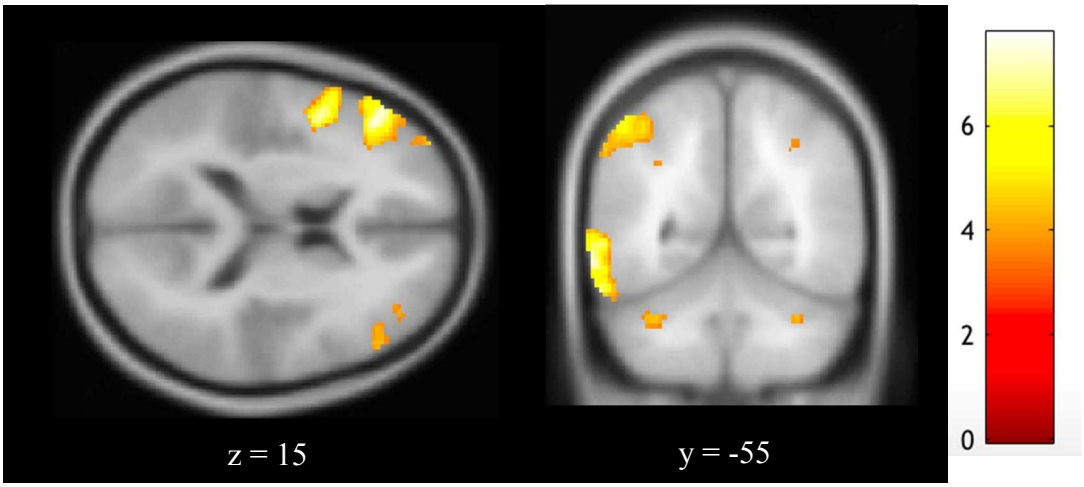

Figure 8. BOLD activity associated with control relative to making identity change judgments for all targets. Heat map refers to t values. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 206.

Table 6
Identification of BOLD Signal Increases for Control Trials Relative to All Identity Change Trials

Neural Region (MNI Coordinates)	*x*	*y*	*z*	No. of voxels	Peak T
Right Cerebellum	30	-60	-32	206	7.26
Left IFG / Lateral PFC	-50	8	18	3129	7.18
Left Parietal / Occipital Lobe	-34	-74	44	1235	7.08
Left Superior Temporal Gyrus	-64	-56	-8	895	6.84
Presupplementary Motor Area	-8	10	52	623	6.6
Right IFG / Lateral PFC	42	6	42	482	5.83
Left Cerebellum	-38	-62	-28	234	5.52
Right Anterior PFC / VLPFC / OFC	30	62	-6	467	5.52
Right IFG / Lateral PFC	48	34	32	428	4.86

Note. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 206. IFG= Inferior frontal gyrus, PFC = prefrontal cortex, VLPFC = ventrolateral prefrontal cortex, OFC = orbitofrontal cortex

Self vs. Friend. One of the primary analyses for this study focused on clarifying the mechanisms underlying the effect of target reported in Study 3. In line with previous work, it was hypothesized that thinking about oneself would recruit more ventral regions of the medial prefrontal cortex, whereas thinking about a friend would recruit more dorsal regions of the medial prefrontal cortex. Statistical activation maps for self trials were compared relative to friend trials. Somewhat surprisingly, brain regions activated for this comparison were somewhat sparse and limited to lateral regions of the brain, including the left lateral posterior parietal cortex and left dorsolateral prefrontal cortex. (Figure 9). Peaks for these activated regions are listed in Table 7.

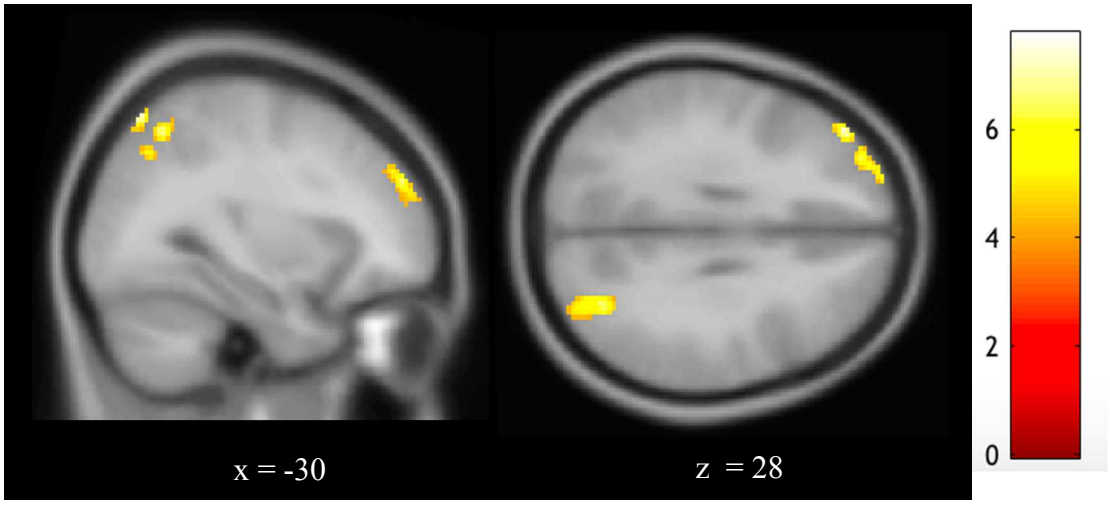

Figure 9. BOLD activity associated with making identity change judgments for self relative to friend. Heat map refers to t values. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent

Table 7
Identification of BOLD Signal Increases for Identity Change Trials for Self Relative to Identity Change Trials for Friend

Neural Region (MNI Coordinates)	*x*	*y*	*z*	No. of voxels	Peak T
Left Precuneus/ PPC /Lateral Parietal	-28	-72	58	300	6.53
Cerebellum	-4	-72	-18	222	6.04
Left dlPFC	-42	42	28	277	6.03
Right IPS / Parietal Cortex	28	-64	58	202	5.8
Fusiform Gyrus / Inferior Temporal	-50	-66	-14	537	5.79
Right IPS / Parietal Cortex	32	-78	34	361	5.72
Left Intraparietal / Inferior Parietal	-54	-42	54	165	5.65

Note. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 158. dlPFC = dorsolateral prefrontal cortex, IPS = Intraparietal Sulcus

Friend vs. Self. Perhaps even more surprisingly, brain regions activated for friend trials relative to self trials included certain cortical midline structure of interest, including the posterior cingulate, as well as the ventral region of the medial prefrontal cortex, extending into the orbitofrontal cortex. Moreover, a region of the right anterior temporal lobe was recruited for this comparison (Figure 10). Peaks for these activated regions are listed in Table 8.

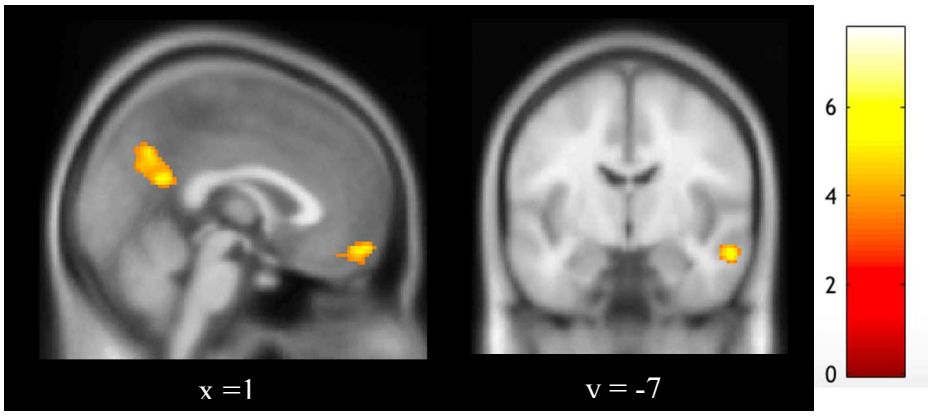

Figure 10. BOLD activity associated with making identity change judgments for friend relative to self. Heat map refers to t values. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 158.

Table 8
Identification of BOLD Signal Increases for Identity Change Trials for Friend Relative to Identity Change Trials for Self

Neural Region (MNI Coordinates)	*x*	*y*	*z*	No. of voxels	Peak T
Posterior Cingulate Cortex	6	-54	20	807	7.76
Anterior Temporal Lobe / Parahippocampal Gyrus	58	-4	-20	221	6.47
vmPFC / OFC	2	60	-14	162	5.53

Note. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 158. vmPFC = ventromedial prefrontal cortex, OFC = orbitofrontal cortex

Self vs. Control. Implicit to the results reported above (person vs. control; self vs. friend) is the assumption that self and friends trials, to a large extent, recruited overlapping activity in cortical midline structure of the brain. However, to verify that assumption, the peaks for each condition (self and friend) relative to control are reported below. Brain regions activated for self trials relative to control trials included certain cortical midline structure of interest, including the posterior cingulate, as well as the medial prefrontal cortex (Figure 11). Peaks for these activated regions are listed in Table 9.

Friend vs. Control. Moreover, as expected, brain regions activated for friend trials relative to control trials included certain cortical midline structure of interest, including the precuneus / posterior cingulate, as well as the medial prefrontal cortex. However, activity here was located more ventrally, extending into the orbitofrontal cortex (Figure 12). Peaks for these activated regions are listed in Table 10.

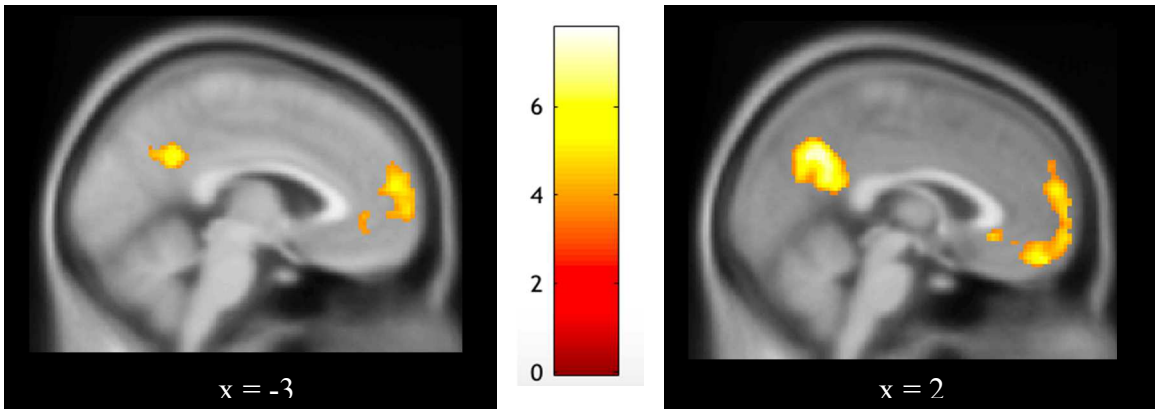

Figure 11. BOLD activity associated with making identity change judgments for self relative to control trials. Heat map refers to t values. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 100.

Figure 12. BOLD activity associated with making identity change judgments for friend relative to control trials. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 100. Heat map refers to t values.

Table 9
Identification of BOLD Signal Increases for Identity Change Trials for Self Relative to Control

Neural Region (MNI Coordinates)	*x*	*y*	*z*	No. of voxels	Peak T
Right Occipital Lobe	34	-84	0	1312	7.21
Posterior Cingulate	-8	-52	30	413	6.28
Medial Prefrontal Cortex	-2	58	16	526	5.25

Note. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 100.

Table 10
Identification of BOLD Signal Increases for Identity Change Trials for Friend Relative to Control

Neural Region (MNI Coordinates)	*x*	*y*	*z*	No. of voxels	Peak T
Precuneus / Posterior Cingulate	4	-56	34	1274	7.56
vmPFC / OFC	4	50	-18	929	6.34
Right Occipital Cortex	24	-100	1	105	6.02
Right Occipital Cortex	34	-72	-16	113	4.93

Note. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 100.

Instruction period. For exploratory purposes, activity recruited during the instruction period of the task for person versus control trials and self versus friend trials was investigated, and it was hypothesized that this preparatory period would recruit similar brain regions as those recruited during the cue periods reported above. However, brain regions recruited for person versus control instructions periods were limited to bilateral regions of the occipital cortex (Figure 13). Peaks for these activated regions are listed in Table 11. No brain regions survived thresholding for the control relative to person-centered instructional cues.

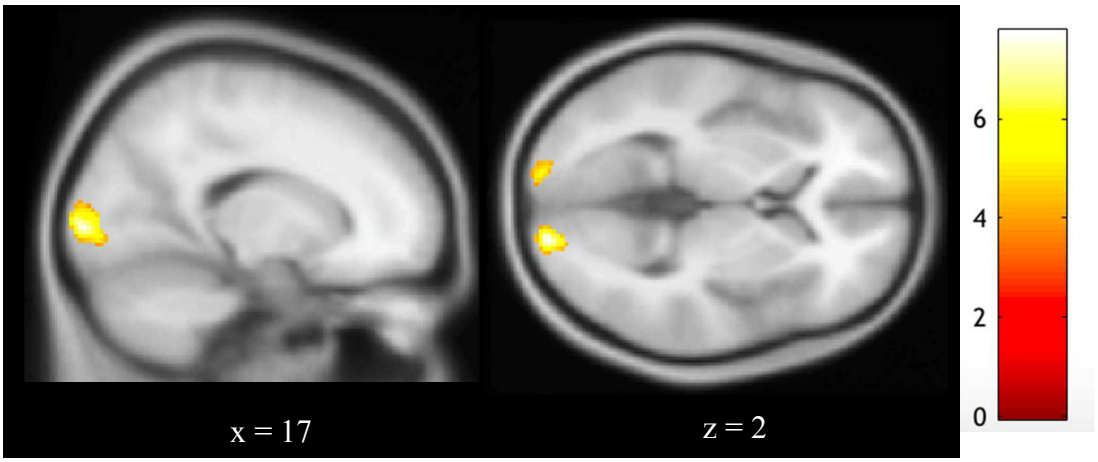

Figure 13. BOLD activity associated with identity change instruction cues for all targets relative to control instruction cues. Heat map refers to t values. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 200.

Table 11
Identification of BOLD Signal Increases for All Person-Related Instruction Cues Relative to Control Instruction Cues

Neural Region (MNI Coordinates)	*x*	*y*	*z*	No. of voxels	Peak T
Left Occipital Cortex	-18	-88	-4	492	6.78
Right Occipital Cortex	18	-96	2	292	6.62

Note. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 200.

Moreover, whereas no brain regions survived thresholding for the self instruction cue relative to the friend instruction cue, brain regions active for the friend instruction cue relative to the self instruction cue included a cluster in right occipital /visual cortex, as well as in left somatosensory cortex (Figure 14), implying some sort of extra visual preparation for friend trials. Peaks for these activated regions are listed in Table 12.

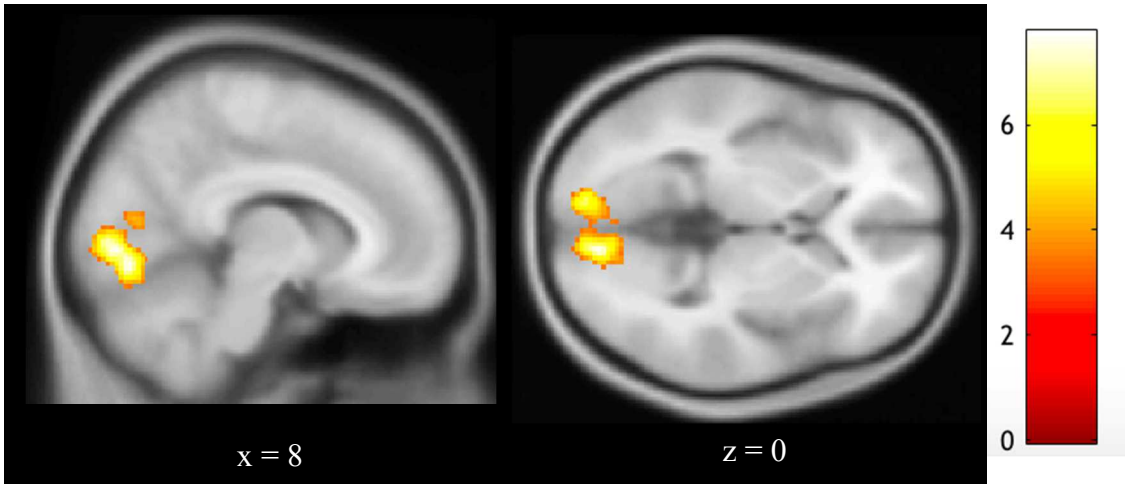

Figure 14. BOLD activity associated with identity change instruction cues for friend relative to self. Heat map refers to t values. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 219.

Table 12
Identification of BOLD Signal Increases for Friend Instruction Cues Relative to Self Instruction Cues

Neural Region (MNI Coordinates)	*x*	*y*	*z*	No. of voxels	Peak T
(Right) Occipital Cortex	12	-78	-6	1498	7.78
Somatosensory Cortex	-42	-30	58	220	4.78

Main Effect of Trait.

Moral vs. Non-Moral Traits. For the effect of trait, it was hypothesized that moral traits, given their relative importance to the self, would recruit more activity in

value and reward-based regions of the brain compared to non-moral traits. Surprisingly, however, no brain regions for this contrast survived thresholding.

Non-Moral vs. Moral Traits. In contrast, non-moral traits relative to moral traits activated a relatively diffuse array of frontal and prefrontal brain regions, including multiple clusters within the left dorsolateral prefrontal cortex / inferior frontal gyrus, left temporal lobe, and bilateral inferior parietal lobe. This contrast also included regions within the cortical midline structures, including the posterior cingulate and the ventromedial prefrontal cortex (Figure 15). Peaks for these activated regions are listed in Table 13.

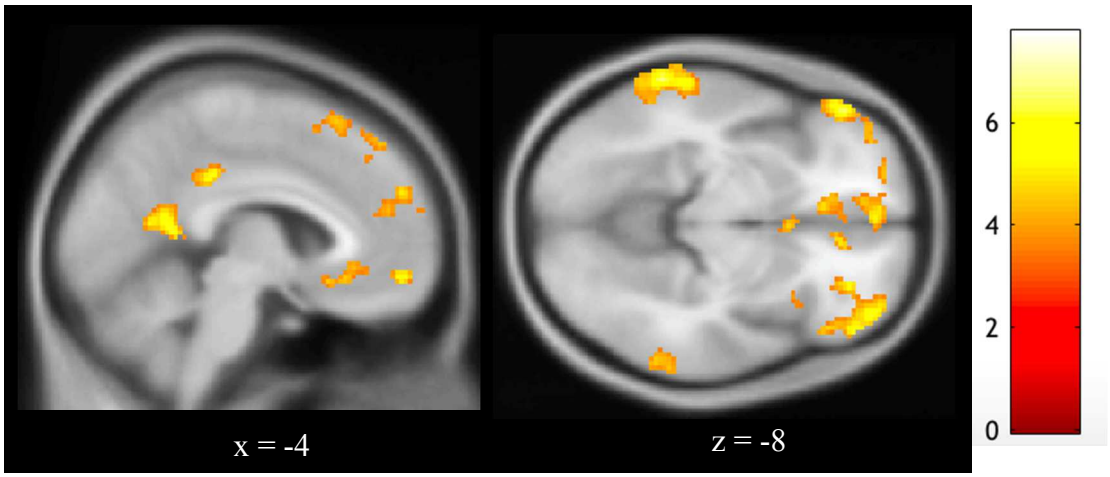

Figure 15. BOLD activity associated with making identity change judgments for non-moral relative to moral traits. Heat map refers to t values. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 205.

Table 13
Identification of BOLD Signal Increases for Non-Moral Identity Change Trials Relative to Moral Identity Change Trials

Neural Region (MNI Coordinates)	*x*	*y*	*z*	No. of voxels	Peak T
Left dlPFC	-22	22	42	2014	7.51
Left Temporal Lobe	-64	-44	-4	1612	7.38
Right Inferior Parietal Lobe	40	-62	34	542	6.74
Left Inferior Parietal Lobe	-38	-74	46	931	6.72
Left Cerebellum	-48	-68	-42	212	6.48
Right Temporal Lobe	58	-28	-24	825	6.35
Left Inferior Frontal Gyrus	-50	40	-8	403	6.22
Posterior Cingulate	-4	-54	14	331	4.7
Right Lateral PFC / OFC	42	52	-8	804	4.68
vmPFC	-4	56	-10	446	5.76
Subgenual Cingulate	-6	32	-4	446	4.88
Posterior Cingulate	-4	-34	36	264	5.59

Note. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 205. dlPFC = dorsolateral prefrontal cortex ; PFC = prefrontal cortex, OFC = orbitofrontal cortex; vmPFC = ventromedial prefrontal cortex

Non-moral vs. moral traits for self. Simple contrasts for non-moral traits relative to moral traits were run for the self and friend condition were examined to determine whether either condition was driving the effect of trait. Many of the regions overlap with those from the overall non-moral versus moral trait contrast, including the bilateral temporal lobe and left dorsolateral prefrontal cortex / inferior frontal gyrus. However, this contrast also revealed activity in dorsomedial prefrontal cortex, a region that was also curiously found in the effect of target for self when viewed at a lower threshold (Figure 16). Peaks for these activated regions are listed in Table 14.

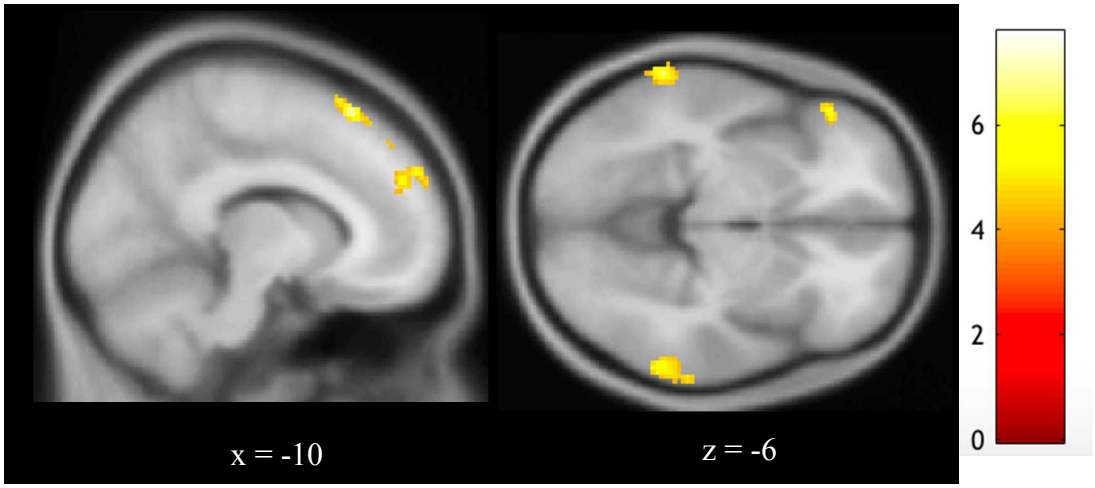

Figure 16. BOLD activity associated with making identity change judgments for self on non-moral relative to moral traits. Heat map refers to t values. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 121.

Table 14
Identification of BOLD Signal Increases for Non-moral Identity Change Trials Relative to Moral Identity Change Trials When Thinking of Self

Neural Region (MNI Coordinates)	*x*	*y*	*z*	No. of voxels	Peak T
dmPFC	-8	28	58	202	6.37
Left Temporal Lobe	-68	-44	-4	779	5.87
Right Inferior Parietal Lobe	44	-60	38	237	5.84
Right Temporal Lobe	60	-32	-18	305	5.41
Left IFG / PFC	-52	28	8	121	5.03
Left Angular Gyrus	-38	-72	44	468	4.95
Medial Superior Frontal Lobe	-8	52	28	129	4.62

Note. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 121. dmPFC = dorsomedial prefrontal cortex, IPL = inference parietal lobe, IFG = inferior frontal gyrus, PFC = prefrontal cortex

Non-moral vs. moral traits for friend. Simple contrasts of moral traits relative to non-moral traits for the friend condition also did not reveal any brain regions surviving threshold, but non-moral traits relative to moral traits for the friend revealed activity in right parahippocampal gyrus / anterior temporal lobe (also found in the effect of trait) and

in ventromedial prefrontal cortex / orbitofrontal cortex (also found in the effect of target) (Figure 17). Peaks for these activated regions are listed in Table 15.

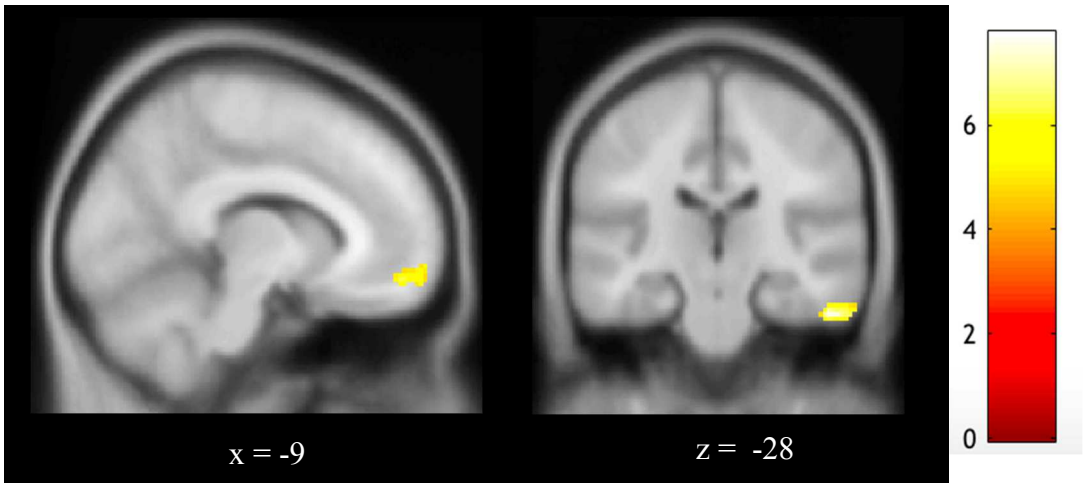

Figure 17. BOLD activity associated with making identity change judgments for friend on non-moral relative to moral traits. Heat map refers to t values. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 113.

Table 15
Identification of BOLD Signal Increases for Non-moral Identity Change Trials Relative to Moral Identity Change Trials When Thinking of Friend

Neural Region (MNI Coordinates)	*x*	*y*	*z*	No. of voxels	Peak T
Parahippocampal Gyrus	54	-20	-28	223	5.7
vmPFC / OFC	-6	56	-12	171	5.36

Note. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 211. vmPFC = ventromedial prefrontal cortex ; OFC = orbitofrontal cortex.

Moral trials vs. control. Implicit to the results reported above (moral vs. non-moral trials) is the assumption that moral and non-moral trait trials, to a large extent, recruited overlapping activity in cortical midline structure of the brain. However, to verify that assumption, the contrasts for each relative to control are reported below. Brain regions activated for moral trials relative to control trials included certain cortical midline

structure of interest, including the posterior cingulate, as well as the medial prefrontal cortex (Figure 18). Peaks for these activated regions are listed in Table 16.

Non-moral trials vs. control. Moreover, as expected, brain regions activated for non-moral trials relative to control trials also included certain cortical midline structure of interest, including the precuneus / posterior cingulate, subgenual cingulate, and medial prefrontal cortex. However, activity in the medial prefrontal cortex for this contrast was slightly more extensive (Figure 19). Peaks for these activated regions are listed in Table 17.

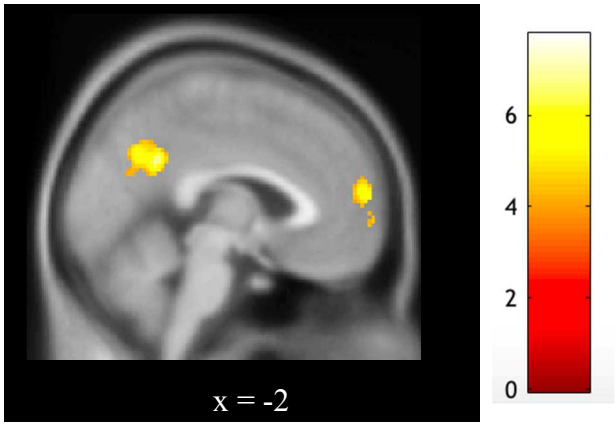

Figure 18: BOLD activity associated with making identity change judgments for moral traits relative to control. Heat map refers to t values. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 100.

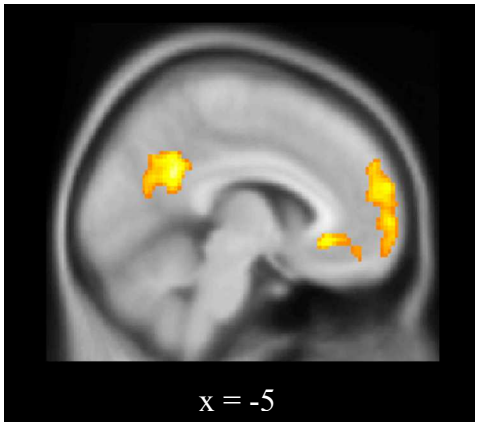

Figure 19: BOLD activity associated with making identity change judgments for non-moral traits relative to control. Heat map refers to t values. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 100.

Table 16
Identification of BOLD Signal Increases for Moral Identity Change Trials Relative to Control

Neural Region (MNI Coordinates)	*x*	*y*	*z*	No. of voxels	Peak T
Right Occipital Cortex	26	-100	12	1030	5.07
Posterior Cingulate	-6	-52	32	620	5.01
Medial Prefrontal Cortex	-2	60	14	172	4.98

Note. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 100.

Table 17
Identification of BOLD Signal Increases for Non-Moral Identity Change Trials Relative to Control

Neural Region (MNI Coordinates)	*x*	*y*	*z*	No. of voxels	Peak T
Precuneus	4	-58	40	1154	7.5
Right Occipital Cortex	24	-98	14	149	6.14
Medial Prefrontal Cortex	0	60	18	1239	5.98
Subgenual Cingulate	-6	30	-6	1239	5.85
Anterior Middle Temporal Lobe	-66	-16	-18	102	5.82
Right Occipital Cortex	28	-80	-8	154	4.22

Note. Correction for multiple comparisons (*FWE P < .05*) applied using threshold of *P < .001* and extent threshold of *k* = 100.

Moral vs. non-moral traits at baseline. Notably, no brain regions survived thresholding in either direction when comparing moral and non-moral traits in the control condition (no identity change judgments). The results reported above were only observed during identity change trials.

Interaction. Based on the behavioral results from Study 3, no strong effect of interaction was hypothesized, but the results are reported here for completeness. No brain regions survived thresholding for the interaction in either condition (moral versus non-moral trait ratings for self relative to non-moral versus moral trait ratings for friend or

non-moral versus moral trait ratings for self relative to moral versus non-moral trait ratings for friend).

Discussion

Main effect of target. I expected that both self and friend trials would reveal activation in cortical midline structure, with trials about the self recruiting more activity in ventral regions of the medial prefrontal cortex and trials about a friend (other) recruiting more activity in dorsal regions of the medial prefrontal cortex. Trials that involved thinking about people in general (relative to the control condition) did invoke activity in these cortical midline structures. However, no brain areas of interest were recruited to a greater extent during self versus friend trials, and, counter to the hypothesis, regions of the precuneus and the ventromedial prefrontal cortex, bordering on orbitofrontal cortex, were more strongly recruited for friend versus self trials. Overall, findings indicate that both self and other trials recruited overlapping activity within the medial prefrontal cortex, with some extra activity recruited for the friend condition in ventral regions.

The results are somewhat surprising given the established neuroscience literature on self and other processing. Self-processing usually recruits regions more ventral of the medial prefrontal cortex, and other-processing usually recruits regions more dorsal of the medial prefrontal cortex. Of course, there is a chance that strong overlapping activity was found because the friend was perhaps too similar to the self. However, previous studies still find significant self versus other differences when using a similar friend as a target (Moore et al., 2014). Moreover, there is a chance that the hypothetical nature of the thought experiment led to highly overlapping activity, but previous studies that used a

hypothetical counterfactual thinking manipulation for self and other showed the traditional separation between their respective patterns within the medial prefrontal cortex.

Instead, the overlap I observed may be indicative of the fact that thinking about personal identity simply does not recruit the same types of target-specific thinking as traditional social psychological phenomena. Rather, given its third-person perspective, thinking about personal identity may be more akin to thinking about personhood more broadly. Alternatively, it is possible that thinking about personal identity recruits a blended set of distinct cognitive processes: thinking both about importance to the self when thinking about the other target and considering others' perceptions of oneself when thinking about the self, requiring complicated social cognitive processes in both conditions.

The increased activity in vmPFC (OFC) for thinking about a friend may be informative for teasing apart underlying mechanisms. Given that this ventral region of the medial prefrontal cortex is typically recruited more for self- vs. other-referential processing, this result may provide evidence that importance to self drives many of the effects for thinking about a target: people care about others changing because it impacts them. Of course, this reasoning is still somewhat speculative. To further probe this hypothesis, future analysis may want to consider using more nuanced pattern analysis techniques to see whether neural patterns for self (vs. friend) on the localizer task are predictive of patterns of activation invoked while thinking about identity change for friend.

However, it is also worth noting that the increased ventromedial activity for friend relative to self (and control) is located quite ventrally, bordering on orbitofrontal cortex, introducing a new host of potential explanations for the finding. Activity in orbitofrontal cortex is less traditionally associated with self-referential thought and moreso with emotion-based decision-making (Bechara, Damasio, & Damasio, 2000). On the one hand, activity in this region might suggest an emotional response to the thought of a friend changing relative to self, an idea explored earlier in this paper: an identity change to a friend will likely impact the subjective experience of the perceiver. On the other hand, activity in this region might suggest some sort of increased deliberation or decision-making, appealing to more of an effort argument (Rushworth, Behrens, Rudebeck, & Walton, 2007). This hypothesis is, to some extent, supported by the fact that thinking about a friend (relative to self) recruits preparatory activity in visual cortex, indicating that more effortful imagination might be required for this condition. Future analysis may want to further investigate this hypothesis by analyzing reaction times for the self and friend conditions. If the other condition is, in fact, more effortful, slower reaction times would be predicted, a behavioral result that has emerged in traditional self and other processing (e.g., Kelley et al., 2002).

Main effect of trait. It was hypothesized that moral traits would recruit more value-based regions of the brain (e.g., vmPFC) relative to non-moral traits. Somewhat surprisingly, thinking about change in moral traits did not recruit different brain activation than thinking about change in non-moral traits. However, thinking about identity change for non-moral traits relative to moral traits recruited a suite of brain regions, including cortical midline regions (precuneus, vmPFC), as well as other regions

related to control-based processing (e.g., bilateral lateral PFC) and semantic processing (e.g., bilateral temporal lobes). Thinking about change in moral and non-moral traits both recruited cortical midline structures relative to control, although, again, these effects were strong for non-moral traits. No differences were seen between moral and non-moral traits at control.

The lack of increased activity for moral trials is surprising given the growing evidence suggesting that moral content does uniquely recruit certain regions of the brain, including the vmFPC, likely for value-based reasons, and sometimes the dmPFC, likely for social cognition reasons). This is not to say that moral trials did not recruit these regions at all while making identity change judgments – in fact, they did, but the activity was stronger for non-moral traits, as well as more diffuse. It is not entirely clear why non-moral traits showed this pattern. Given that non-moral traits were rated higher in competence, it is possible that this "extra" processing is required for reasoning about more skill-based identity changes.

However, it is also possible that considering these non-moral traits simply required more effort than reasoning about moral traits; moral traits may have been easier to process. If moral traits are, in fact, essential to a person, then it seems plausible that making judgments about how these moral traits might change would be more intuitive and less effortful. The claim is, to some degree, supported by the fact that although differences are observed for thinking about non-moral and moral traits during judgments of identity change, no differences are seen between thinking about these same moral and non-moral traits during control, suggesting that there is something special about thinking about morality in the context of an individual. In other words, there may not be anything

special about moral traits, per se, but moreso something special about the effects that those moral traits have on oneself or another individual; moral traits may only be contextually special. Future analysis could investigate the degree to which response times to trials for moral and non-moral trait words differ to further probe the possibility of an appeal to effort.

Interaction. No interactions between target and trait were hypothesized, and none were observed. The null result was unsurprising given the behavioral results observed in Study 3. Moreover, a significant interaction would have been difficult to interpret given the possibility of inflated effects of interactions under some conditions within neuroimaging (Chavez & Wagner, in prep). Of course, it is possible that there are situations in which an interaction would be present – a case in which moral traits might be less important to identity than other traits for a given participant - but such a case lies outside the scope of this paper.

Limitations and Future Directions

As mentioned earlier, one potential limitation of the current study is that targets were instructed to think about a relatively close friend. Although participants were explicitly instructed that the friend should *not* be their best friend, they were instructed that the selected friend should be someone they know well and with whom they interact on a daily basis. This degree of closeness increases the probability that effects between self and friend would not be detected, meaning that, in many ways, this was a very conservative comparison, and the specific nature of the psychological differences between the two are worth considering carefully. Future analysis may want to further examine the influence of target closeness by incorporating self-reported information

about the relationship with the friend (e.g., closeness) in the target analysis (e.g., as a parametric modulator).

The experiment may have also been limited in its use of an event-related design. Behavioral studies to date investigating the role of target (see Study 3) have used a blocked design such that participants make all ratings for one target before making ratings for another target. Instead, trials for targets in this design were not blocked. This event-related design was chosen to both 1) maintain psychological interest on the relatively long trials and 2) to optimize for the ability to investigate trial-level trait effects (e.g., trait ratings) without worrying about signal interference from a blocked design. There is a chance that requiring participants to switch between targets washed out some of possible effects. Previous studies have shown strong self-other neural asymmetries using similar event-related designs, but this type of design may be less amenable to hypothetical thought experiments.

The implications of the results of the main effect of target are that the traditional findings for processing self versus other information, assumed to be robust, may be limited to thinking about the self in particular ways – ones that might be idiosyncratic to the way psychologists think about the concept of self. However, the self is a vast topic, long considered from different perspectives in other disciplines. Despite the overlap in content (the self), many of these competing perspectives have, to date, been overlooked within the fields of psychology and neuroscience. Considering and incorporating them may be important for extending and challenging traditionally held assumptions within both fields. To continue to push and understand the boundary conditions underlying certain psychological effects, future work will need to continue to consider different

perspectives on the self. For example, future studies might consider additional hypothetical and narrative approaches to self that rely upon a complicated area of underlying cognitive processes.

Another limitation of the present study is that it used valence-matched trait words in the tightest possible condition. Although this tightly controlled comparison was a strength for interpreting any possible neural differences, it did not allow as much room for detecting the moral self effect itself. In traditional studies on the moral self-effect, moral traits are compared to a large swath of person-level characteristics, including memories, non-valence matched personality traits, preferences, and desires. Future neuroimaging work may want to consider investigating how processing moral trait words compares with processing words in these other categories, albeit considering that other dimensions like valence may significantly impact the results

Given the relative overlap between thinking about identity change for moral and non-moral traits relative to control, it is also possible that any effect of thinking about moral traits is simply not detectable using univariate techniques (e.g., subtracting the average degree of activation during one condition from the average of another). Future approaches may want to consider using multivariate classification techniques between the value-based or the self-based localizer and the identity change task to see whether a trained classifier can predict any differences between moral and non-moral traits. Previous work has demonstrated that trait-based patterns are more detectable using multivariate approaches (e.g., Tamir, Thornton, Contreras, & Mitchell, 2016), and given the person-centered nature of the trait words used in the study, it is plausible that these trait words are both encoded in similar, person-based regions of the brain, but in different

ways depending on their self-relevance or their value-signal. Future analysis may also consider incorporating behavioral ratings of change for a trait (as a parametric modulator). It is possible that although considering moral versus non-moral traits as separate categories did not reveal any neural differences under direct region-by-region comparisons, considering traits that were perceived to lead to more identity change would. Additionally, it is possible that actual perceptions of individual targets might have washed out the moral self effect to some degree, and future analysis should consider adjusting identity change ratings for those perceptions

Based on these results, the degree to which an explanation for morality as a special category should be sought remains unclear. On the one hand, morality may simply just be essential to identity, and there may not be any other explanatory mechanism (e.g., value-based processing or social cognition) required for understanding its uniqueness or its importance. On the other hand, many explanatory mechanisms have been hypothesized in the previous literature, and can be explored using novels experimental designs and methods. The degree to which this information would add explanatory power should continue to be considered in future work.

General Discussion

Knowns

This suite of experiments replicated the original moral self effect (while ruling out valence as an underlying mechanism), revealed a possible effect of target for self versus other, and identified a number of brain regions underlying both effects, albeit in unexpected directions. In particular, overlapping activity was observed for the self and other conditions, with more activity in regions of the ventromedial prefrontal cortex /

orbitofrontal cortex observed when thinking about moral and non-moral trait changes in a close friend. Overlapping activity was also observed for the moral and non-moral conditions, with more overall activity (in cortical midline structures, temporal, and parietal regions) observed for thinking about whether non-moral traits might change.

The goal of the study was to address two separate but related questions. First, the study investigated whether thinking about personal identity for different targets is similar to engaging in traditional person perception, and the answer, for the most part, seems to be no. Although the behavioral evidence demonstrates that thinking about self yields larger perceived identity change than thinking about a friend, the neural evidence suggests that the mechanisms underlying judgments of identity change for oneself and a friend do not differ in expected ways. Thinking about identity change recruited brain regions typically involved in person perception, more broadly, but the typical ventral to dorsal split for thinking about self and other, respectively, was not observed. Rather, there was strong overlap when thinking about personal identity for self and other, suggesting that thinking about personal identity is not the same as engaging in traditional person perception, which typically results in strong differences between self and other.

Second, the study aimed to better understand why morality is special. Rather than obtaining explanatory answers, however, this study obtained evidence to the contrary: that morality may not actually be as special as has been hypothesized in the literature to date. Although the behavioral evidence replicated the finding that changes to morality are perceived to lead to more identity change than changes to other personality traits, the neural evidence did not support this claim. Rather, there was stronger activity in a variety of brain regions when thinking about non-moral versus moral traits whereas no such

activity was found for thinking about moral versus non-moral traits, suggesting that there may not be any particular mechanism driving the privileged status of moral traits within respect to personal identity.

Unknowns & Future Directions

In the face of these conclusions, it is important to note that these results are complicated and represent an important and broader move towards a more naturalistic neuroscience (Zaki & Ochsner, 2009; Schonberg, Fox, & Poldrack, 2011; Tikka & Kaipainen, 2014). The thought experiments included in this study are abstract and require the detection of subtle and nuanced differences between conditions. As a result, a complicated array of underlying neural mechanisms were hypothesized. Humans are not simple, nor are the contents of our thoughts or experiences. As neuroscience moves to apply its findings to real-world contexts (e.g., Gabrieli, Ghosh, & Whitfield-Gabrieli, 2015), understanding the ways in which people process more naturalistic, complicated sets of stimuli becomes increasingly important.

Of course, the stimulus sets used in this thought experiment are not directly naturalistic. Engaging in arm chair philosophy, at least on its surface, is quite different from the everyday experience of most people. As such, the conclusions born from this experiment are primarily limited to those speculating about the nature of personal identity and of morality. However, engaging in hypothetical thought experiments about one's own identity change or that of a close, actual friend could easily be adapted to have important implications for a variety of translational work. For example, clarifying how people process hypothetical identity changes in another individual, whether they are suffering from disease or undergoing a gender transition, can help to understand and

predict the ways that we treat other individuals; clarifying how we process hypothetical identity changes in ourselves, whether we are motivated to become a different person through addiction treatment or whether we fear letting go of the person we are now whilst trying to lose weight, can help to understand and predict the ways in which individuals pursue their own goals. In this sense, this suite of studies sets a foundation for considering a complicated array of processes that lie at the core of who we are.

CHAPTER IV

DISCUSSION

One aim of the current dissertation project was to contribute toward an integrative study of self. An integrative study of self can be defined as one that unifies seemingly separate conceptions of self to meet broader functionalist and ontological needs. Chapter 2 examined elements of self-complexity through an empirical investigation on multiples selves, and Chapter 3 examined elements of self-continuity through an empirical investigation on personal identity and the moral self. Both projects contribute to an integrative study of self in distinct but meaningful ways.

The work on multiple selves (Chapter 2) is integrative in that it applies network analysis techniques to the study of self. Network analysis techniques are only beginning to make their way into the field of psychology, more broadly (Costantini et al., 2014), and the set of studies presented in Chapter 2 are the first to use this network-based approach to examine the structure of the self. This introduction of network-based methods to the study of the self has implications for both advancing intervention-based work, as well as for addressing bigger questions about the nature of the self. Specifically, through measuring more nuanced about the relationships between different self-aspects, the network-based approach provides important information for designing more targeted identity-based interventions. In turn, more targeted identity-based interventions will allow researchers to avoid some of the uncertainties regarding the mechanisms underlying current, self-based manipulations (e.g., self-affirmation). This network-based approach also helps to address bigger questions about the nature of the self. One key

question centers around how the self is unified across content (Klein 2012b). This network-based approach provides a new method for measuring the degree of integration (or unification) across self-content.

Not only is the work on multiple selves integrative, but the work on personal identity and morality (Chapter 3) is integrative in that it uses neuroscience methods and social-psychological theory to addresses age-old questions in philosophy. Specifically, the study examines first and third-person judgments of identity change to determine whether traditional findings within neuroscience and social psychology apply to thinking about personal identity, which traditionally only considers a third-person perspective on the self. The findings of this integrative approach have a number of implications for both advancing intervention-based work, as well as for addressing bigger questions about the nature of the self. Notably, the study found that morals lead to more perceived identity change for both self and other, suggesting that morality constitutes an essential part of an individual's identity. As a result, future identity-based interventions may want to consider targeting an individual's moral identity in order to elicit long-lasting change. Moreover, the study found that, to a large degree, thinking about identity change for self and for other recruits overlapping brain activity. These findings suggest that thinking about personal identity for self and for other may be different than thinking about standard self and other-based perceptions, an insight which helps to inform bigger questions about how the self is unified across time (Klein, 2012b).

Together, then, both of these studies consider broader definitions of self that have the potential to inform more translational work. A common theme between the definitions of self that were considered in these studies is that they both, to some degree,

draw on social elements of the self. The self-aspects that people list in the study on multiple selves tend to be related to social roles (Chapter 2), and morality is hypothesized to be essential to an individual's personal identity because of its social nature (Chapter 3). These ideas are, in large part, in accordance with the "looking glass" perspective on the self, which hypothesizes that self-perceptions are drawn from the ways that other people in our social world see us (Cooley, 1902; Mead, 1934; Yeung & Martin, 2003). Indeed, evidence suggests that meta-perceptions, or perceptions of how others perceive us, serve as a particularly important source of self-knowledge (Carlson, Furr, & Vazire, 2010; Carlson & Furr, 2009), and that perceptions of others' acceptance or rejection of us heavily determines self-esteem (Leary, Tambor, Terdal, & Downs, 1995), suggesting that social input largely influences both the information and the affect associated with self-perception. Given the largely social nature of the self, future work may want to consider using certain social identities as an inlet for encouraging identity change. For example, individuals may be more amenable to identity change if the targeted identity is associated with an important social group (e.g., a favorite sports team) or if the intervention occurs during a period of social transition (e.g., moving to a new city or ending a romantic relationship).

In addition to taking an integrative perspective on the self that considers broader definitions of the self for informing future translational work, another aim of the current dissertation was to contribute towards a larger organizational framework that considers the self, not only across traits, but also across roles (contexts) and narratives (time). Both sets studies reported in Chapters 2 and 3 contribute to this larger organizational framework at both levels. The multiple selves study (Chapter 2) contributes to the

literature on roles in that it characterizes the relationships between the different roles that people identify to be most important in their lives. Through investigating the structure of these relationships, researchers gain the ability to better understand the goals and motivates that guide an individual across different contexts. Moreover, the study contributes to the literate on narrative in that it characterizes the collection of characters that represent meaningful aspects of an individual's life. Through investigating the ways in which these characters integrate and come together, researchers gain the ability to understand the types of narrative relationships that, over time, give an individual meaning and purpose.

The study on personal identity and morality (Chapter 3), too, contributes to the larger organizational framework that considers both roles and narratives. First, in identifying morality as an especially core component of an individual's own, self-perceived identity, the study suggests that morality may also comprise a particularly salient motivator or role for an individual. Notably, however, the types of self-aspects listed by most individuals do not tend to be moral in nature (McConnell et al., 2011), raising the question of whether moral content is core to identity in the same way that particular self-aspects and roles are. Future work should consider exploring how the essential nature of morality actually plays out in everyday roles and contexts. The study also contributes to the broader literature on narrative. Although the study of narrative has been distinguished from the study of personal identity (Peacocke, 2014), there is an important temporal component that unites these approaches. Through studying perceptions of identity change, the study contributes to that element of narrative that identifies core elements of an individual's identity across time.

Despite their contributions toward this broader, overall self-framework, the studies were limited in similar ways. Specifically, despite the push to move towards studying more role and narrative-based elements of self, both studies were still drawn from and inspired by trait-based approaches. The study on multiple selves investigated the relationships between the different self-aspects by asking participants to rate each of their self-aspects on a number of trait dimensions; the study on personal identity and morality investigated perceptions of identity change by asking participants to imagine different degrees of trait-based change. In this sense, the trait-based approaches used in these studies may have distracted from the more contextual and narrative elements included in the studies. However, given that trait words form the basis for the study of identity (McAdams, 1995), a trait approach, to some degree, is required in order to study of self at these higher levels. The trait-based approaches used here also allow for building incrementally off of previous work. Even so, future work will want to consider new methods and approaches for moving away from such trait-heavy approaches and towards more naturalistic ones.

Similarly, both are studies are also limited, to some extent, in that they both replicate and build off of previous paradigms. Replication, of course, is valuable for the advancement of knowledge, more broadly (Maxwell, Lau, & Howard, 2015). However, sticking to established paradigms too strictly may limit the ability to move towards a more naturalistic study of self.

Future work will want to consider applying more contextualized and naturalistic approaches towards the study of role-based and narrative-based elements of self. A broader movement within social psychology is pushing for the study of real-world, actual

human behavior (Baumeister, Vohs, & Funder, 2007), and methods are being developed for doing so. For example, the electronic audio recorder (EAR) has been an effective tool for sampling real-world behaviors (Mehl, Pennebaker, Crow, Dabbs, & Price, 2001). Similar techniques could be advantageous for examining real-world self-integration (or complexity) across contexts as well as for sampling the types of stories that people tell about themselves.

Taking a more contextualized and naturalistic approach is, by nature of the method, much more difficult to do in neuroscience. However, a push for more naturalistic study is being made there, as well (Zaki & Ochsner, 2009; Schonberg, Fox, & Poldrack, 2011; Tikka & Kaipainen, 2014). For example, future studies might want to consider examining the neural mechanisms underlying processing of actual self-narratives. Narratives could be collected from participants using the life story interview (McAdams, 2008), which asks people to detail high points, low points, and turning points from the course of their life thus far. Participants could then be presented with "scenes" from the written transcription of their life story while in the scanner. Although researchers are starting to understand mechanisms underlying story-processing, more generally (Baldassano, Chen, Zadbood, Pillow, Hasson, & Norman, 2016; Chen, Leong, Honey, Yong, Norman, & Hasson, 2017), and although the neural mechanisms underlying trait-based self-relevant processing is well-understood (Denny, Kober, Wagner, & Ochsner, 2012; Wagner, Haxby, & Heatherton, 2012), much less is known about the mechanism underlying the processing of self-narrative. The stability of identity provides a source of value, and thus value-based mechanisms would be hypothesized to play a larger role for more stable identities. Moreover, because event-structure is generally encoded as shifts in

the patterns of certain neural activity, pattern-based classification techniques would be particularly useful. Analyses of interest could probe content about particular self-aspects, investigating, for example, how stories about a goal-self or a moral-self are represented, and perhaps whether these patterns are predictive of identity consistent real-world behaviors, informing the degree to which narrative-based information (versus other types of self-relevant information) plays a role in determining future behavior. Moreover, these narratives could be probed across time, investigating, for example, how narratives changes after certain transformative life events, informing bigger questions about mechanism underlying identity change. Investigation of self-narrative is feasible and provides a viable future direction for work in this area.

Although definitions of self within the field vary, studies such as that proposed for investigating the neural mechanisms underlying self-narrative can help to provide a landscape for understanding the relationships between the different elements of self. These definitions can be refined by work in philosophy and can inform translational approaches that seek to understand the effects of acting upon and engaging different elements of the self. In this sense, the current work is simply the start of a broader move towards an integrative study of self.

APPENDIX A

TRAIT WORDS USED TO RATE SELF-ASPECTS

Positive:	Negative:
Capable	Disagreeing
Comfortable	Disorganized
Communicative	Hopeless
Confident	Immature
Energetic	Incompetent
Friendly	Indecisive
Fun and Entertaining	Inferior
Giving	Insecure
Happy	Irresponsible
Hardworking	Irritable
Independent	Isolated
Intelligent	Lazy
Interested	Like a failure
Lovable	Sad and Blue
Mature	Self-centered
Needed	Tense
Optimistic	Uncomfortable
Organized	Unloved
Outgoing	Weary
Successful	Worthless

APPENDIX B

EXAMPLE RESPONSES TO WRITING PROMPTS

Manipulation (Self-Integration) Condition:

Self-Aspects: Electrician, Wife, Commuter, Dog Owner

"Commuter self wants to get to work on time to make sure electrician self keeps a job. Commuter wants to be timely so electrician doesn't miss anything on the first part of the job. Electrician self wants to bring home money so that wife self can contribute to the household. Wife self wants to make the family as happy as possible which dog owner self can help with by keeping the dogs healthy, happy and well-groomed. Dog owner self does her part so wife self can do other chores such as laundry and vacuuming. Electrician self also earns money to help dog owner self afford the care of the dogs. Electrician self provides the money for gas and car so that commuter self can commute. Commuter self helps wife self by calling the husband during the daily commute. Wife self and dog owner self like to joke with the husband about the dogs. This makes every one feel good and happy. Wife self makes sure lunch is packed for everyone and that helps commuter self get out the door in a timely fashion. Dog owner self says goodbye to the dogs and helps commuter feel better about going to work."

Control Condition:

"I wake up early in the morning. As usual I struggle to get up from bed, but I do, I put some clothes on, get my food and go out. I walk to work where I meet my colleagues. We chat for sometime and then everyone goes to work on their sections. I usually stay there during the two smaller breaks that we have. While I'm there I'm thinking of how to make more money while listening to something sports related. During lunch break I meet up with my colleges and my boss and we grab something to eat. After that I do the same thing until it's end of work day. I get off work and start planning on what should I do next. I go to the gym and work out, then if I need something from the groceries store I go and get it. I walk back to my place while meeting the usual people who do the same at that time. When I'm home I shower and browse the internet until it's time to go to bed. Usually I talk to some people from home on social media and/or watch a tv show too. Before going to bed I think about what should I prepare for the next day and go to bed."

APPENDIX C

MORAL AND NON-MORAL TRAIT WORDS

Moral Traits	Non-Moral Traits
Truthful	Intelligent
Trustworthy	Happy
Kind	Smart
Genuine	Capable
Empathic	Knowledgeable
Principled	Confident
Sincere	Independent
Loyal	Easygoing
Compassionate	Determined
Fair	Charismatic
Tolerant	Realistic
Responsible	Funny
Understanding	Educated
Selfless	Skillful
Merciful	Productive
Faithful	Perceptive
Forgiving	Articulate
Respectful	Adaptable
Considerate	Resourceful
Altruistic	Punctual

REFERENCES CITED

Alfano, M. (2015). How one becomes what one is called: On the relation between traits and trait-terms in Nietzsche. *Journal of Nietzsche Studies, 46*(2), 261–269.

Anderson, N. H. (1968). Likableness ratings of 555 personality- trait words. *Journal of Personality and Social Psychology, 9*, 272 – 279.

Aquino, K., & Reed II, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423.

Ariel, B. (2012). Deterrence and moral persuasion effects on corporate tax compliance: findings from a randomized controlled trial. *Criminology*, *50*(1), 27-69.

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*(4), 596.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2016). Discovering event structure in continuous narrative perception and memory. *Neuron, 95,* 709- 721.

Baumeister, R. F. (1998). The self. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of Social Psychology*, *1*(4), 680–740.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior?. *Perspectives on Psychological Science*, *2*(4), 396-403.

Bechtel, W. (2008). *Mental mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. Taylor & Francis.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561–571.

Bench, S. W., Schlegel, R. J., Davis, W. E., & Vess, M. (2015). Thinking about change in the self and others: The role of self-discovery metaphors and the true self. *Social Cognition*, *33*(3), 169–185.

Bench-Capon, T., & Modgil, S. (2017). Norms and value based reasoning: justifying compliance and violation. *Artificial Intelligence and Law*, *25*(1), 29-64.

Bergner, R. M. (2017). What is a person? What is the self? Formulations for a science of psychology. *Journal of Theoretical and Philosophical Psychology, 37*(2), 77.

Berkman, E.T. & Falk, E.B. (2013). Beyond brain mapping: Using neural measures to predict real-world outcomes. *Current Directions in Psychological Science, 22*, 45-50.

Berkman, E. T., Hutcherson, C. A., Livingston, J. L., Kahn, L. E., & Inzlicht, M. (2017). Self-control as value-based choice. *Current Directions in Psychological Science*, *26*(5), 422-428.

Berkman, E.T., Kahn, L.E., & Livingston, J.L. (2016). Valuation as a mechanism of self-control and ego depletion. In E.R. Hirt (Ed.), *Self-Regulation and Ego Control* (pp. 255-279). New York: Elsevier.

Berkman, E. T., Livingston, J. L., & Kahn, L. E. (2017). Finding the "self" in self-regulation: The identity-value model. *Psychological Inquiry*, *28*(2-3), 77-98.

Berridge, K. C. (1996). Food reward: brain substrates of wanting and liking. *Neuroscience & Biobehavioral Reviews*, *20*(1), 1-25.

Black, J. E., & Reynolds, W. M. (2016). Development, reliability, and validity of the Moral Identity Questionnaire. *Personality and Individual Differences*, *97*, 120-129.

Blok, S., Newman, G., & Rips, L. J. (2005). Individuals and their concepts. *Categorization Inside and Outside the Lab*, 127-149.

Blumenthal, M., Christian, C., Slemrod, J., & Smith, M. G. (2001). Do normative appeals affect tax compliance? Evidence from a controlled experiment in Minnesota. *National Tax Journal*, 125-138.

Bolderdijk, J. W., Steg, L., Geller, E. S., Lehman, P. K., & Postmes, T. (2013). Comparing the effectiveness of monetary versus moral motives in environmental campaigning. *Nature Climate Change*, *3*(4), 413.

Bondy, J. A., & Murty, U. S. R. (1976). *Graph Theory With Applications* (Vol. 290). London: Macmillan.

Brown, C. M., Bailey, V. S., Stoll, H., & McConnell, A. R. (2016). Between two selves: Comparing global and local predictors of speed of switching between self-aspects. *Self and Identity*, *15*(1), 72-89.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977-990.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network. *Annals of the New York Academy of Sciences*, *1124*(1), 1-38.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, *11*(2), 49-57.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186.\

Campbell, J. I., & Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavior research methods*, *44*(4), 1255-1265.

Campbell, J. D., Trapnell, P. D., Heine, S. J., Katz, I. M., Lavallee, L. F., & Lehman, D. R. (1996). Self-concept clarity: Measurement, personality correlates, and cultural boundaries. *Journal of Personality and Social Psychology, 70*(1), 141-156.

Carlson, E. N., & Furr, R. M. (2009). Differential meta-accuracy: People understand the different impressions they make. *Psychological Science, 20,* 1033-1039.

Carlson, E. N., Furr, R. M., & Vazire, S. (2010). Do we know the first impressions we make? Evidence for idiographic meta-accuracy and calibration of first impressions. *Social Psychological and Personality Science, 1,* 94-98.

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology*, *56*, 453-484.

Chavez, R. S., Heatherton, T. F., & Wagner, D. D. (2016). Neural population decoding reveals the intrinsic positivity of the self. *Cerebral Cortex*, *27*(11), 5222-5229.

Chavez, R.S. & Wagner, D.D. (preprint) Mass univariate testing biases the detection of interaction effects in whole-brain analysis of variance.

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*(1), 115-125.

Chiu, C. Y., Hong, Y. Y., & Dweck, C. S. (1997). Lay dispositionism and implicit theories of personality. *Journal of Personality and Social Psychology*, *73*(1), 19.

Cohen, S., & Hoberman, H. M. (1983). Positive events and social supports as buffers of life change stress. *Journal of Applied Social Psychology, 13*, 99–125.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24*, 385–396.

Cooley, C. H. (1902) 1983. *Human Nature and the Social Order.* Transaction.

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, *54*, 13-29.

Craik, F. I., Moroz, T. M., Moscovitch, M., Stuss, D. T., Winocur, G., Tulving, E., & Kapur, S. (1999). In search of the self: A positron emission tomography study. *Psychological Science*, *10*(1), 26-34.

D'Argembeau, A., Stawarczyk, D., Majerus, S., Collette, F., Van der Linden, M., & Salmon, E. (2010). Modulation of medial prefrontal and inferior parietal cortices when thinking about past, present, and future selves. *Social Neuroscience*, *5*(2), 187–200. http://doi.org/10.1080/17470910903233562

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4, Pt. 1), 377.

De Brigard, F., Spreng, R. N., Mitchell, J. P., & Schacter, D. L. (2015). Neural activity associated with self, other, and object-based counterfactual thinking. *NeuroImage*, *109*(C), 12–26. http://doi.org/10.1016/j.neuroimage.2014.12.075

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. New York, NY: Plenum.

Dennett, D. C. (1992). The self as a center of narrative gravity. In *Self and consciousness: Multiple perspectives*. Hillsdale, NJ: Erlbaum.

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self-and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of cognitive Neuroscience*, *24*(8), 1742-1752.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*, 48(4), 1-18. http://www.jstatsoft.org/v48/i04/.

Erskine, R. G., & Trautmann, R. L. (1993). The process of integrative psychotherapy. In *The boardwalk papers: Selections from the 1993 eastern regional transactional analysis association conference* (pp. 1-26). Madison, WI: Omnipress.

Ersner-Hershfield, H., Wimmer, G. E., & Knutson, B. (2008). Saving for the future self: Neural measures of future self-continuity predict temporal discounting. *Social Cognitive and Affective Neuroscience*, *4*(1), 85-92.

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Aliaga, A. E., ... & Oya, H. (2018). FMRIPrep: a robust preprocessing pipeline for functional MRI. *bioRxiv*, 306951.

Everett, J. A. C., Skorburg, J. A., Livingston, J. L., Chituc, V. & Crockett, M. J. (in prep). How moral is the "moral self?" Morality, community, and identity persistence.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160.

Festinger, L. (1957). A theory of cognitive dissonance. Evanston, IL: Row Peterson.

Fossati, P., Hevenor, S. J., Lepage, M., Graham, S. J., Grady, C., Keightley, M. L., ... & Mayberg, H. (2004). Distributed self in episodic memory: neural correlates of successful retrieval of self-encoded positive and negative personality traits. *Neuroimage*, *22*(4), 1596-1604.

Frank, L. E., & Nagel, S. K. (2017). Addiction and moralization: the role of the underlying model of addiction. *Neuroethics*, *10*(1), 129-139.

Frankfurt, H. G. (1988). Freedom of the Will and the Concept of a Person. In *What Is a Person?* (pp. 127-144). Humana Press.

Fujita, K., Trope, Y., Liberman, N., & Levin-Sagi, M. (2006). Construal levels and self-control. *Journal of Personality and Social Psychology*, *90*(3), 351.

Funder, D. C. & Doboroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology, 52*, 409-418.

Gabrieli, J. D., Ghosh, S. S., & Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron*, *85*(1), 11-26.

Garfield, J. L., Nichols, S., Rai, A. K., & Strohminger, N. (2015). Ego, egoism and the impact of religion on ethical experience: What a paradoxical consequence of buddhist culture tells us about moral psychology. *The Journal of Ethics*, *19*(3-4), 293-304.

Glisky, E. L., & Marquine, M. J. (2009). Semantic and self-referential processing of positive and negative trait adjectives in older adults. *Memory*, *17*(2), 144-157.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of personality and social psychology*, *106*(1), 148.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., ... & Handwerker, D. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*, 160044.

Hackel, L. M., & Zaki, J. (2017). My Brain Contains Multitudes: The Value of a Flexible Approach to Identity. *Psychological Inquiry*, *28*(2-3), 99-102.

Han, H. (2017). Neural correlates of moral sensitivity and moral judgment associated with brain circuitries of selfhood: a meta-analysis. *Journal of Moral Education*, *46*(2), 97-113.

Hardy, S. A., & Carlo, G. (2005). Identity as a source of moral motivation. *Human Development*, *48*(4), 232-256.

Hardy, S. A., & Carlo, G. (2011). Moral identity: What is it, how does it develop, and is it linked to moral action?. *Child Development Perspectives*, *5*(3), 212-218.

Hare, R.M., 1952, *The Language of Morals*, New York: Oxford University Press.

Harman, G. (1999, January). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. In *Proceedings of the Aristotelian Society* (pp. 315-331). Aristotelian Society.

Hayes, S. C. (2004). Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. *Behavior Therapy*, *35*(4), 639-665.

Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive Science*, *41*(3), 744-767.

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review*, *94*, 319-340.

Higgins, E. T., Roney, C. J., Crowe, E., & Hymes, C. (1994). Ideal versus ought predilections for approach and avoidance distinct self-regulatory systems. *Journal of Personality and Social Psychology*, *66*(2), 276.

Hopper, J. R., & Nielsen, J. M. (1991). Recycling as altruistic behavior: Normative and behavioral strategies to expand participation in a community recycling program. *Environment and Behavior*, *23*(2), 195-220.

Hume, D. (1978). A Treatise of Human Nature, with text revised and notes by PH Nidditch. *Oxford: The Clarendon Press.(First Edition 1888)*.

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, *87*(2), 451–462.

James, W. (1890). *The Principles of Psychology*.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality*, *61*(4), 521-551.

Johnson, S. C., Baxter, L. C., Wilder, L. S., Pipe, J. G., Heiserman, J. E., & Prigatano, G. P. (2002). Neural correlates of self-reflection. *Brain*, *125*(8), 1808-1814.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, *10*(12), 1625–1633.

Katzko, M. W. (2003). Unity versus multiplicity: A conceptual analysis of the term "self" and its use in personality theories. *Journal of Personality*, *71*(1), 83-114.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience*, *14*(5), 785-794.

Kelley, W. M., Wagner, D. D., & Heatherton, T. F. (2015). In search of a human self-regulation system. *Annual Review of Neuroscience*, *38*, 389-411.

Kemp, G. (2005). Philosophy as fiction: Self, deception, and knowledge in Proust. *Philosophy and Literature, 29*(2), 498-500.

Kim, K., & Johnson, M. K. (2015). Activity in ventromedial prefrontal cortex during self-related processing: Positive subjective value or personal significance? *Social Cognitive and Affective Neuroscience, 10*(4), 494– 500.

Klein, S. B. (2012a). The self and its brain. *Social Cognition*, *30*(4), 474.

Klein, S.B. (2012b). Self, memory, and the self-reference effect: An examination of conceptual and methodological issues. *Personality and Social Psychology Review*, *16*(3), 283-300.

Klein, W. M., & Harris, P. R. (2009). Self-affirmation enhances attentional bias toward threatening components of a persuasive message. *Psychological Science*, *20*(12), 1463-1467.

Kross, E. & Ayduk, O. (2017). Self-distancing: Theory, research and current directions. In J. Olson & M. Zanna (Eds.), *Advances in Experimental Social Psychology, 55*, 81-136.

Kross, E., Bruehlman-Senecal, E., Park, J., Burson, A., Dougherty, A., Shablack, H., et al. (2014). Self-talk as a regulatory mechanism: How you do it matters. *Journal of Personality and Social Psychology*, *106*(2), 304–324.

Landy, J. (2004). Proust, his narrator, and the importance of the distinction. *Poetics Today*, *25*(1), 91-135.

Leary, M. R., Tambor, E. S., Terdal, S. K., & Downs, D. L. (1995). Self-esteem as an interpersonal monitor: The sociometer hypothesis. *Journal of Personality & Social Psychology, 68*(3), 518-530.

Leary, M. R., Tate, E. B., Adams, C. E., Batts Allen, A., & Hancock, J. (2007). Self-compassion and reactions to unpleasant self-relevant events: the implications of treating oneself kindly. *Journal of Personality and Social Psychology*, *92*(5), 887.

Lempert, K. M., & Kable, J. W. (2017). Separating identity and value in the identity-value model. *Psychological Inquiry*, *28*(2-3), 103-107.

Levine, M., & Perkins, D. V. (1980, August). Tailor making life events scale. Paper presented at the meeting of the American Psychological Association, Montreal.

Lin, W. J., Horner, A. J., & Burgess, N. (2016). Ventromedial prefrontal cortex, adding value to autobiographical memories. *Scientific Reports*, *6*, 28630.

Linville, P. W. (1987). Self-complexity as a cognitive buffer against stress-related illness and depression. *Journal of Personality and Social Psychology, 52*, 663–676.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433-442.

Livingston, J. L., Kahn, L. E., & Berkman, E. T. (in prep). The identity-value model: elaborations, applications, and future directions. *Psychological Inquiry, 28,* 157-164.

Livingston, J. L., Skorburg, J. A., Everett, J., A. C., Ferguson, M. A., De Brigard, F., Sinnott-Armstrong, W., & Karns, C. (in prep). Finding the "moral" self? An event-related fMRI study.

Locke, J. (1690/2009). An essay concerning human understanding. New York, NY: WLC Books.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*(2), 224.

Markus, H., & Kunda, Z. (1986). Stability and malleability of the self-concept. *Journal of Personality and Social Psychology, 51*(4), 858–866.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?. *American Psychologist*, *70*(6), 487.

McAdams, D. P. (1995). What do we know when we know a person? *Journal of Personality*, *63*(3), 365-396.

McAdams, D. P. (2008). *The Life Story Interview*. Retrieved from https://www.sesp.northwestern.edu/docs/Interviewrevised95.pdf

McAdams, D. P. (2013). The psychological self as actor, agent, and author. *Perspectives on Psychological Science*, *8*(3), 272-295.

McAdams, D. P., & Guo, J. (2015). Narrating the generative life. *Psychological Science*, *26*(4), 475-483.

McConnell, A. R. (2011). The multiple self-aspects framework: Self-concept representation and its implications. *Personality and Social Psychology Review*, *15*(1), 3-27.

McConnell, A. R., Renaud, J. M., Dean, K. K., Green, S. P., Lamoreaux, M. J., Hall, C. E., & Rydell, R. J. (2005). Whose self is it anyway? Self-aspect control moderates the relation between self-complexity and well-being. *Journal of Experimental Social Psychology*, *41*(1), 1-18.

McConnell, A. R., Rydell, R. J., & Brown, C. M. (2009). On the experience of self-relevant feedback: How self-concept organization influences affective responses and self-evaluations. *Journal of Experimental Social Psychology*, *45*(4), 695-707.

McConnellL, A. R., & Strain, L. M. (2007). Content and structure of the self-concept. *The Self*, 51-73.

Mead, G. H. (1934). *Mind, Self, and Society: From the Standpoint of a Social Behaviorist*, edited by Charles W. Morris. University of Chicago Press.

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, *33*(4), 517-523.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *17*(8), 1306-1315.

Molden, D. C., Hall, A., Hui, C. M., & Scholer, A. A. (2017). Understanding how identity and value motivate self-regulation is necessary but not sufficient: A motivated effort-allocation perspective. *Psychological Inquiry*, *28*(2-3), 113-121.

Moore, W. E. III (2015). *Sharing all the way to the bank: A neuroimaging investigation of differential self-disclosure*. Retrieved from University of Oregon Dissertation Database.

Moore, W. E. III, Merchant, J. S., Kahn, L. E., & Pfeifer, J. H. (2014). 'Like me?': Ventromedial prefrontal cortex is sensitive to both personal relevance and self-similarity during social comparisons. *Social Cognitive and Affective Neuroscience, 9*, 421-426.

Mumford, J. A., & Nichols, T. E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage*, *39*(1), 261-268.

Neff, K., D., & Seppala, E. (2016). Compassion, Well-Being, and the Hypoegoic Self. In K. W. Brown & M. Leary (Eds*), Oxford Handbook of Hypo-egoic Phenomena: Theory and Research on the Quiet Ego (pp. 189 -202)*. Oxford University Press.

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin*, *40*(2), 203-216.

Nichols, S., & Bruno, M. (2010). Intuitions about personal identity: An empirical study. *Philosophical Psychology*, *23*(3), 293-312.

Northoff, G. (2017). Personal identity and Cortical Midline Structures (CMS) – Do temporal features of CMS neural activity transform into "self-continuity?" *Psychological Inquiry, 28*(2-3), 122-131.

Northoff, G., & Hayes, D. J. (2011). Is our self nothing but reward? *Biological Psychiatry, 69*(11), 1019–1025.

Olson, E. T. (2016). Personal identity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016 ed.). Retrieved from http://plato .stanford.edu/archives/spr2016/entries/identity-personal/;10.1002/9781118922590.ch7

Packer, D. J., & Cunningham, W. A. (2009). Neural correlates of reflection on goal states: the role of regulatory focus and temporal distance. *Social Neuroscience*, *4*(5), 412-425.

Parfit, D. (1971). Personal identity. *The Philosophical Review*, *80*(1), 3-27.

Peacocke, C. (2014). *The mirror of the world: subjects, consciousness, and self-consciousness*. Oxford University Press.

Peters, M. L., Flink, I. K., Boersma, K., & Linton, S. J. (2010). Manipulating optimism: Can imagining a best possible self be used to increase positive future expectancies?. *The Journal of Positive Psychology*, *5*(3), 204-211.

Pfeifer, J. H., Masten, C. L., Borofsky, L. A., Dapretto, M., Fuligni, A. J., & Lieberman, M. D. (2009). Neural correlates of direct and reflected self-appraisals in adolescents and adults: When social perspective taking informs self-perception. *Child Development, 80*, 1016-1038.

Pfeifer, J. H., Kahn, L. E., Merchant, J. S., Peake, S. J., Kim Veroude, Masten, C. L., et al. (2013). Longitudinal change in the neural bases of adolescent social self-evaluations: Effects of age and pubertal development. *Journal of Neuroscience, 33*(17), 7415–7419.

Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience*, *27*(37), 9984-9988.

Quoidbach, J., Gilbert, D. T., & Wilson, T. D. (2013). The end of history illusion. *Science*, *339*(6115), 96-98.

Rafaeli-Mor, E., & Steinberg, J. (2002). Self-complexity and well- being: A review and research synthesis. *Personality and Social Psychology Review, 6*, 31–58.

Ramarajan, L. (2014). Past, present and future research on multiple identities: Toward an intrapersonal network approach. *The Academy of Management Annals*, *8*(1), 589-659.

Renaud, J. M., & McConnell, A. R. (2002). Organization of the self-concept and the suppression of self-relevant thoughts. *Journal of Experimental Social Psychology*, *38*(1), 79-86.

Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin*, *143*(2), 117.

Roberts, B. W., & Donahue, E. M. (1994). One personality, multiple selves: Integrating personality and social roles. *Journal of Personality*, *62*(2), 199-218.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rosenburg, M. (1979). *Conceiving the self*. Basic Books.

Ross, L. (1977). The Intuitive Psychologist And His Shortcomings: Distortions in the Attribution Process. In *Advances in Experimental Social Psychology* (Vol. 10, pp. 173-220). Academic Press.

Ross, L. D., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of Personality and Social Psychology*, *35*(7), 485.

Rozin, P., & Singh, L. (1999). The moralization of cigarette smoking in the United States. *Journal of Consumer Psychology*, *8*(3), 321-337.

Rushworth, M. F. S., Behrens, T. E. J., Rudebeck, P. H., & Walton, M. E. (2007). Contrasting roles for cingulate and orbitofrontal cortex in decisions and social behaviour. *Trends in Cognitive Sciences*, *11*(4), 168-176.

Salsman, J. M., Lai, J. S., Hendrie, H. C., Butt, Z., Zill, N., Pilkonis, P. A., ... & Cella, D. (2014). Assessing psychological well-being: self-report instruments for the NIH Toolbox. *Quality of Life Research*, *23*(1), 205-215.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: remembering, imagining, and the brain. *Neuron*, *76*(4), 677- 694.

Schechtman, M. (1996). The Constitution of Selves. Ithaca. *NY: Cornell University*

Schmeichel, B. J., & Vohs, K. (2009). Self-affirmation and self-control: affirming core values counteracts ego depletion. *Journal of Personality and Social Psychology*, *96*(4), 770.

Schonberg, T., Fox, C. R., & Poldrack, R. A. (2011). Mind the gap: bridging economic and naturalistic risk-taking with cognitive neuroscience. *Trends in Cognitive Sciences*, *15*(1), 11-19.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: another look at the availability heuristic. *Journal of Personality and Social Psychology*, *61*(2), 195.

Scott, W. A. (1969). Structure of natural cognitions. *Journal of Personality and Social Psychology, 12*, 261–278.

Sedikides, C., Gaertner, L., & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of Personality and Social Psychology*, *84*(1), 60.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667–677.

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: Dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience*, *34*(13), 4741–4749.

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. *Advances in Experimental Social Psychology*, *38*, 183-242.

Showers, C. (1992). Compartmentalization of positive and negative self-knowledge: Keeping bad apples out of the bunch. *Journal of Personality and Social Psychology*, *62*(6), 1036.

Smith, S. H., & Cohen, L. H. (1993). Self-complexity and reactions to a relationship breakup. *Journal of Social and Clinical Psychology, 12*, 367–384.

Sklar, A. Y., & Fujita, K. (2017). On When and How Identity Value Impacts Self-Control Decisions. *Psychological Inquiry*, *28*(2-3), 153-156.

Slingerland, E., & Collard, M. (Eds.). (2011). *Creating consilience: Integrating the sciences and the humanities*. Oxford University Press.

Sripada, C. (2016). Self-expression: A deep self theory of moral responsibility. *Philosophical Studies*, *173*(5), 1203-1232.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in Experimental Social Psychology*, *21*(2), 261-302.

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*, *12*(4), 551-560.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition, 131*, 159–171.

Strohminger, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science*, *26*(9), 1469-1479.

Tamir, D. I., & Mitchell, J. P. (2011). The default network distinguishes construals of proximal versus distal events. *Journal of Cognitive Neuroscience*, *23*(10), 2945-2955.

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences*, *113*(1), 194–199. http://doi.org/10.1073/pnas.1511905112

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193.

Tikka, P., & Kaipainen, M. Y. (2014). From naturalistic neuroscience to modeling radical embodiment with narrative enactive systems. *Frontiers in Human Neuroscience*, *8*, 794.

Tobia, K. P. (2016). Personal identity, direction of change, and neuroethics. *Neuroethics*, *9*(1), 37–43.

VandenBos, G. R. (2007). *APA dictionary of psychology*. Washington, DC: American Psychological Association.

Vanderlinden, J., Van Dyck, R., Vandereycken, W., Vertommen, H., & Jan Verkes, R. (1993). The dissociation questionnaire (DIS-Q): Development and characteristics of a new self-report questionnaire. *Clinical Psychology & Psychotherapy*, *1*(1), 21-27.

Vazire, S., & Carlson, E. N. (2010). Self-knowledge of personality: Do people know themselves? *Social and Personality Psychology Compass*, *4*, 605-620.

Wager, T. D., & Nichols, T. E. (2003). Optimization of experimental design in fMRI: a general framework using a genetic algorithm. *Neuroimage*, *18*(2), 293-309.

Wagner, D. D., Haxby, J. V., & Heatherton, T. F. (2012). The representation of self and person knowledge in the medial prefrontal cortex. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(4), 451–470.

West, D. B. (2001). *Introduction to Graph Theory* (Vol. 2). Upper Saddle River: Prentice Hall.

White, R. E., Prager, E. O., Schaefer, C., Kross, E., Duckworth, A. L., & Carlson, S. M. (2016). The "batman effect": Improving perseverance in young children. *Child Development*. http://doi.org/10.1111/cdev.12695

Wittgenstein, L. (1953). *Philosophical Investigations*. New York, NY: Macmillan.

Yankouskaya, A., Humphreys, G., Stolte, M., Stokes, M., Moradi, Z., & Sui, J. (2017). An anterior–posterior axis within the ventromedial prefrontal cortex separates self and reward. *Social Cognitive and Affective Neuroscience*, *12*(12), 1859-1868.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuro- imaging data. *Nature Methods, 8*(8), 665–670.

Yeung, K. T., & Martin, J. L. (2003). The looking glass self: An empirical test and elaboration. *Social Forces*, *81*(3), 843-879.

Zaki, J., & Ochsner, K. (2009). The need for a cognitive neuroscience of naturalistic social cognition. *Annals of the New York Academy of Sciences*, *1167*(1), 16-30.

Zhu, Y., Zhang, L., Fan, J., & Han, S. (2007). Neural basis of cultural influence on self-representation. *NeuroImage, 34*(3), 1310–1316.